Crystallographic studies on alcohol dehydrogenase from *Drosophila*

by

Elspeth Jane Gordon

Ph.D. Thesis

University of Edinburgh 1993

Declaration

I declare that this thesis was composed by me, that the work of which it is a record was done by me, except where stated in the thesis. This work has not been accepted elsewhere in any previous application for a degree. All of the sources of information have been acknowledged.

As we advance in life we learn the limits of our abilities

– Froude

Short studies on great subjects

To my family

Acknowledgements

I would like to thank my supervisor, Dr. Lindsay Sawyer, for help and support during this project.

I would also like to thank the following people:

- Prof. Roser González-Duarte and Dr. Silvia Atrian, for providing purified alcohol dehydrogenase enzyme from both *Drosophila lebanonensis* and *Drosophila melanogaster.*

- Stella Bury for producing the original crystal form of the *Drosophila lebanonensis* alcohol dehydrogenase crystals.

- Dr's Miroslav Papiz, Pierre Rizkallah and Sean McSweeney for help during data collection and processing. Also, for many helpful discussions.

- Prof. Neil Isaacs and Dr. Andy Freer for allowing us to use their Xentronics area detector.

- Prof. Simon Phillips for allowing us to use the Xentronics area detector. Also, Dr. Colin Groom for help with data processing and Dr. Nobutoshi Ito for many helpful e-mail messages.

- Dr. Deb Ghosh for providing us with the polyalanine model of 3α, 20β-hydroxysteroid dehydrogenase.

- Eleanor Dodson for help with self rotation function analysis and heavy atom refinement.

- Dr. Andrew Leslie for help with processing image plate data.

- Dr. Lluis Ribas de Pouplana for many useful discussion.

Finally, I would like to thank the people in the lab who were there at all stages of the project with help and advice. Dr. Paul Taylor for help and guidance,

especially with computing problems. Dr. Paul Adams for help and discussions. Also, Joao, Mary, Alan and Fred for being around.

Thanks also go to Fred, Alan, Linda, Joao and Sean for proof reading bits and pieces.

Abstract

Alcohol dehydrogenases are enzymes that catalyse the oxidation of alcohols to aldehydes and ketones. Alcohol dehydrogenase from the fruit-fly *Drosophila*, is a particularly efficient enzyme for this reaction and it is a member of the short chain dehydrogenase protein family. Little structural information has been determined for this family, although extensive biochemical and enzymological studies have been carried out. The aim of this project was to determine the crystal structure of the alcohol dehydrogenase enzyme from *Drosophila*.

Chapter one contains an introduction to dehydrogenases, concentrating mainly on the short chain dehydrogenase family and in particular on the *Drosophila* alcohol dehydrogenase. The short chain family contains more then twenty proteins from a diversity of natural sources, including prokaryotes and humans. The widespread nature of this family indicates its general importance as a protein family, and makes it essential to produce a structural profile for it. The structure of alcohol dehydrogenase from *Drosophila* together with biochemical data, may be used to validate the reaction mechanisms that have already been proposed. The structure of alcohol dehydrogenase from *Drosophila* will also provide an unique opportunity for comparisons to be made between a short chain and a medium chain alcohol dehydrogenase. Such a comparison will give insight into the evolutionary origin of the different alcohol dehydrogenase protein types, and may suggest why more than one protein family has evolved to carry out a single detoxification function.

Alcohol dehydrogenase from *Drosophila* has been crystallized and two crystal forms have been observed. Most crystals were plate-like (form A) and only 0.05 mm in their shortest dimension. Form A crystals diffract X-rays weakly and initial crystal characterization was carried out using the Synchrotron Radiation Source, Daresbury, U.K. Twinning was a severe problem with this crystal form. The second crystal form (form B) was grown in the absence of NAD^+ and with DTT added to all crystallization buffers. This form is more suitable for X-ray diffraction studies since the crystals diffract X-rays to better than 2 Å resolution. Form B crystals are monoclinic with unit cell dimensions, a = 81.24(6), b = 55.75(4), c = 109.60(7)Å and β = 94.26(9)° and they have two dimers in the asymmetric unit. However, it appears that a smaller rotated cell is also valid at low resolution, with unit cell dimensions, a = 70.60, b = 55.75, c = 65.74 Å and β = 106.95° and with one dimer in the asymmetric unit.

Native data and derivative data have been collected on both crystal forms. Most data were collected on form B crystals and analysis and statistics for this data are included. Several data collection systems were used in the course of this study and the relative advantages of each system are discussed.

Attempts to phase the crystallographic data have included both isomorphous replacement and molecular replacement techniques. The molecular replacement study was possible as a result of the recent structural determination of another short chain dehydrogenase, 3α, 20β-hydroxysteroid dehydrogenase. However, only a partially refined polyalanine model of this dehydrogenase was available and this made phase determination by molecular replacement a challenging problem. A preliminary solution has been refined. This molecular replacement solution gives an R-factor = 38.9% for 9237 reflection. The solution was refined using simulated annealing with data ($|F| > 2\sigma$) between 3-15Å.

Two isomorphous derivative data sets have been used to calculate an MIR map. These derivatives are weakly substituted and the strongest site is common to both derivatives. Eight heavy atom positions were refined using maximum likelihood phase refinement. They gave an mean figure of merit 0.43 for 4684 acentric reflections.

Both a molecular replacement map and a MIR map have been calculated (to 3.5Å resolution). The maps were of poor quality and at this stage an unambiguous chain trace is impossible. However, the quality of the maps has been improved by using density modification techniques. A preliminary interpretation of these maps has been made and is discussed.

Table of Contents

List of Figures

List of Tables

Abbreviations

Proteins

ADH : Alcohol dehydrogenase

Adh : Alcohol dehydrogenase gene

DADH : *Drosophila* alcohol dehydrogenase

DlADH : alcohol dehydrogenase from *Drosophila lebanonensis*

DmADH : alcohol dehydrogenase from *Drosophila melanogaster*

HSD : 3α, 20β-hydroxysteroid dehydrogenase

RDH : Ribitol dehydrogenase

SDH : Sorbitol dehydrogenase

HBDH : 2, 3-Dihydro-2, 3-dihydroxybenzoate dehydrogenase

GLU DH : Glucose dehydrogenase

PGDH : 15-Prostaglandin dehydrogenase

LADH : Horse Liver alcohol dehydrogenase

LDH : Lactate dehydrogenase

GAPDH : Glyceraldehyde-6-phosphate dehydrogenase

MDH : Malate dehydrogenase

GDH : Glutamate dehydrogenase

Data sets

NATX : Native data set scaled with XSCALE

NATRA : Native data set scaled with ROTAVATA/AGROVATA

CMN : Data from crystal soaked in 1 mM 2-chloromercuri-4-nitrophenol

GCMN : Data from crystal soaked in 6 mM 2-chloromercuri-4-nitrophenol

PTC : Data from crystal soaked in 0.1 mM potassium platinium chloride

PTC2 : Data from crystal soaked in 0.5 mM potassium platinium chloride

PTC3 : Data from crystal soaked in 1 mM potassium platinium chloride

TAPC : Data from crystal soaked in 5 mM tetrammineplatinium (II) chloride

GHG : Data from crystal soaked in 3 mM mercury chloride

Symbols

K_m : Dissociation constant

V_{max} : Maximum reaction velocity

K_M : Apparent dissociation constant

M : Molarity

Å : Angstrom

r.m.s : Root mean square

M : Molar

E : Energy

r : Distance

a b c : Unit cell dimensions

α, β, γ : Unit cell angles

hkl : Miller indices in X-ray diffraction

F_{hkl} : Complex structure factor in X-ray diffraction

xyz : Fractional atomic coordinates

R : Residual (crystallographic or otherwise)

B : Isotropic temperature factor

F_o : Observed structure factor

F_c : Calculated structure factor

f : Atomic scattering factor

h : Reflection hkl

P_{uvw} : Patterson function

uvw : Fractional coordinates in Patterson space

$\rho(x)$: Electron density at a point x

λ : Wavelength of X-rays

θ : Angle of diffraction

F_{HLE} : Heavy atom lower estimate

F_P : Modulus of protein structure factor

F_{PH} : Modulus of derivative protein structure factor

F_H : Modulus of heavy atom structure factor

$\mathbf{F_P}$: Complex of protein structure factor

$\mathbf{F_{PH}}$: Complex of derivative protein structure factor

$\mathbf{F_H}$: Complex of heavy atom structure factor

K : Scale factor

ΔI : Anomalous intensity difference

w_h : Weighting factor for heavy atom refinement

FOM : Mean figure of merit

m : Figure of merit

D_{iso} : Isomorphous Difference

R_{iso} : Fractional isomorphous difference

α_{best} : Centroid phase

ω, ϕ, κ : Spherical polar coordinates

α, β, γ : Euler angles

$\theta_1, \theta_2, \theta_3$: Lattmann's pseudo Eulerian angles

l, m, n : Direction cosines

Miscellaneous

DTT : Dithiothreitol

PEG : Polyethylene glycol

MPD : 2-methyl-2,4-pentanediol

PDB : Brookhaven Protein Data Bank

FFT : Fast Fourier transform

SA : Simulated annealing

ND : Normal probablilty distrubution

CM-200 : Thinking Machines Corporation Connection Machine 200

SRS : Synchrotron Radiation Source

SERC : Science and Engineering Research Council

NAD^+ : Nicotinamide adenosine dinucleotide

AMP : Adenosine monophosphate

$NADP^+$: Nicotinamide adenosine dinucleotide phosphate

RER : Rapid equilibrium random

Techniques

MR : Molecular Replacement

IR : Isomorphous Replacement

MIR : Multiple Isomorphous Replacement

SIR : Single Isomorphous Replacement

PTF : Phased translation function

c.d. : Circular Dichroism

RF : Rotation function

TF : Translation function

PC : Patterson correlation refinement

DM : Density modification

SF : Solvent flattening

MA : Noncrystallographic averaging

HM : Histogram matching

SE : Sayre's equations

Chapter 1

Introduction

1.1 Alcohol dehydrogenase from *Drosophila*

The *Drosophila* alcohol dehydrogenase gene-enzyme system is a well known biological system which has been studied at all levels, from organism to DNA. The system has been the subject of many reviews that cover a wide range of biological topics:

- biochemistry and population genetics (van Deldon, 1982)
- gene expression (Sofer and Martin, 1987)
- alcohol metabolism (Geer *et al.*, 1990)
- genetics and biochemistry (Chambers, 1988)
- kinetic studies (Winberg and McKinley-McKee, 1992).

Central to all of these studies is the character of the alcohol dehydrogenase gene protein product, the alcohol dehydrogenase enzyme (E.C. 1.1.1.1). A structure/function understanding of the *Drosophila* alcohol dehydrogenase (DADH) enzyme is required for a full understanding and appreciation of the *Drosophila* alcohol dehydrogenase gene (*Adh*) function.

DADH is a member of the short chain dehydrogenases family. The short chain dehydrogenases family was first characterized in 1981 (Jörnvall *et al.*, 1981) when similarities between the alcohol dehydrogenase (ADH) from *Drosophila* and the ribitol dehydrogenase from *Klebsiella* were noted. There are now more than twenty enzymes that have been identified as belonging to the short chain dehydrogenase family and a tertiary structure for one member of this family has recently become available (Ghosh *et al.*, 1991). Little is known about the mechanism of the short chain dehydrogenases but some of the common features and conserved residues have been identified and are discussed below. There have been many biochemical and enzymatic studies carried out on the DADH enzyme

which have allowed two reaction mechanism to be proposed (McKinley-McKee and Winberg, 1992; Ribas de Pouplana and Fothergill-Gilmore, 1993). Confirmation of these mechanisms awaits the determination of a crystal structure. The possible application of the mechanisms to all short chain dehydrogenases can then be explored.

The structure of the DADH will provide a unique opportunity for comparisons to be made between a short chain and a medium chain alcohol dehydrogenase (Eklund *et al.*, 1976; Hurley *et al.*, 1991). This comparison will give an insight into the evolutionary origin of the different alcohol dehydrogenase protein types and may determine why more than one family of enzymes has been evolved to carry out a single detoxification function.

1.2 Alcohol dehydrogenases

Alcohol dehydrogenases (EC.1.1.1.1) constitute a group of enzymes, that catalyse the conversion of alcohols to aldehydes and/or ketones. They use NAD^+ or $NADP^+$ as a cofactor in this reaction. The protein sequences of the ADH enzymes have been classified, and on that basis, four enzyme families have been identified as long chain, medium chain, short chain and iron-dependent dehydrogenases (Persson *et al.*, 1991). There is a great deal of sequence variation within these families but at least one member of each family has ethanol dehydrogenase activity, hence the common nomenclature of alcohol dehydrogenase.

It is not clear whether the different types of dehydrogenases have evolved from a common ancestral gene as an example of divergent evolution or whether they are the result of different proteins converging towards a common function (Brändèn, 1980)

1.2.1 Long chain dehydrogenases

Long chain dehydrogenases have only recently been identified as a dehydrogenase family (Inoue *et al.*, 1989). The only member is a 72 kilodalton alcohol dehydrogenase from *Acetobacter aceti*. The characteristics of this family are not well defined.

1.2.2 Medium chain dehydrogenases

The medium chain family (previously termed the long chain family until the discovery of the alcohol dehydrogenase from *Acetobacter aceti* (Inoue *et al.*, 1989)), is the most well characterized of all of the dehydrogenase families (Rossmann *et al.*, 1975). Much is known about the structures, mechanisms and variability of these enzymes. Each enzyme molecule is composed of several subunits, with each subunit having approximately 350 residues. The members of this family share a sequence identity of about 25% (Jörnvall, 1977) which is very low for functionally related enzymes but crystal structures of members of this family reveal a strongly conserved tertiary structure (Rossmann *et al.*, 1975; Eklund and Brändén, 1987). Each enzyme subunit has two structural/functional domains: a catalytic domain and a coenzyme-binding domain. The structure of the coenzyme-binding domain is conserved throughout the family and is a Rossmann fold (Rossmann, 1974). The Rossmann fold is a common motif in enzymes that bind and use NAD^+ or $NADP^+$ as a cofactor (see section 1.3). In the medium chain family, the coenzyme-binding domain is usually found in the C-terminal end of the protein. However, the lactate dehydrogenase (LDH), malate dehydrogenase (MDH) and glyceraldehyde-3-phosphate dehydrogenase (GAPDH) enzymes have a Rossmann fold in the N-terminal end of the polypeptide chain. The structure of the catalytic domain is not conserved within the medium chain family since this domain is involved in substrate binding and its structure varies according to the function of the dehydrogenase.

A recent publication (Danielsson and Jörnvall, 1992) presents information that suggests that the classical alcohol dehydrogenase (class 1 enzyme) has evolved

from a functional class III form, the glutathione-dependent formaldehyde dehydrogenase line. This implies that the other dehydrogenase families, discussed herein, are more closely related to the glutathione-dependent formaldehyde dehydrogenase than they are to the classical liver alcohol dehydrogenase.

The medium chain dehydrogenases are metalloenzymes and they tend to have one or two zinc atoms per subunit. There are several exceptions to this, for example, the discovery that δ crystallin is a member of the medium chain family, and that this protein has no zinc atom bound. Further, no dehydrogenase activity has yet been shown (Borras *et al.*, 1989). LDH and MDH only have no zinc atom while horse liver alcohol dehydrogenase (LADH) has two. In the LADH, one atom lies in the active site and is liganded by 2 Cys and a His. Variations in this liganding occur in other dehydrogenases e.g. sheep liver sorbitol dehydrogenase (Eklund *et al.*, 1990). The active site zinc atom when present has been shown to be essential for activity as have the active site cysteine residues.

The degree of variation within the medium chain dehydrogenases has been examined and explains the multiplicity of enzymes and isoenzymes (Jörnvall *et al.*, 1990). For the medium chain family there are three regions of the molecule which are prone to variations: the active site region, the subunit interaction region and the area around the second zinc atom. Similar species variations, isozyme-like multiplicities and mutants have been observed for the short chain dehydrogenase family.

1.2.3 Iron-dependent dehydrogenases

The third alcohol dehydrogenase family comprises the iron-dependent dehydrogenases (Scopes, 1983; Williamson and Paquin, 1987). There is little known about the iron-dependant dehydrogenases and currently only two bacterial enzymes have been identified as belonging to this family.

1.2.4 Short chain dehydrogenases

The short chain dehydrogenases include a number of distantly related enzymes from prokaryote, insect, humans and other eukaryotes. Their functions cover the breakdown of alcohols, steroids, sugars and prostaglandins. Little is known about the mechanisms of these enzymes, but structural information is beginning to emerge. Short chain dehydrogenases are composed of a number of subunits with each subunit approximately 250-325 residues long. There are no metal atoms bound to the protein (Schwartz *et al.*, 1976; Thatcher, 1977; Place *et al.*, 1980; Villarroya *et al.*, 1989). Hydrophobicity plots, secondary structure predictions (Thatcher and Sawyer, 1980) and residue distributions have identified a dinucleotide binding domain in the N-terminal end of the protein. This has been proposed for other members of this family (Jörnvall *et al.*, 1984a; Persson *et al.*, 1991) and confirmed by the structure determination of HSD (Ghosh *et al.*, 1991). Sequence identity between members of the short chain family are in the range of 20-25%, with each subunit having 241-327 residues in the polypeptide chain. Like many distantly related proteins glycine is the most conserved residue, thus indicating the importance of steric restrictions and that the short chain dehydrogenases have a common fold; the C-terminus region shows the most sequence variation. Generally the short chain enzymes have few (Schwartz *et al.*, 1976) or no cysteine residues (Marekov *et al.*, 1990). The cysteine residues present in the DADH are not catalytic (Chen *et al.*, 1990).

Alignments of the short chain dehydrogenase sequences (see Figure 1–1) show that there are patterns within the sequences of high similarity, for example, residues 5-15, 130-140, 150-160 and 225-235 (Persson *et al.*, 1991). Overall, there is a variation in the extent of conservation within the internal regions of the proteins and only a few hydrophilic regions are conserved. The overall comparisons suggest the possibility of related mechanisms and domain properties within the short chain family.

The tertiary structure of holo-HSD shows, that unlike the medium chain dehydrogenases this short chain dehydrogenase has no inter-subunit domains. It

```
                          10        20        30        40        50
>DlADH          M-------D--LTNKNVIFVAALGGIGLDTSRELVKRNLKNFVILDRVENPTALAELKAI
>DmADH          M-------SFTLTNKNVIFVAGLGGIGLDTSKELLKRDLKNLVILDRIENPAAIAELKAI
>HSD            M-ND-------LSGKTVIITGGARGLGAEAARQAVAAGAR-VVLADVLDEEGAATA----
>RDH            MKHSVSSMNTSLSGKVAAITGAASGIGLECARTLLGAGAK-VVLIDREGEK----LNKLV
>SDH            M------------NQVAVVIGGGQTLGAFLCHGLAAEGYR-VAVVDIQSDKAANVAQEIN
>GLU DH         MYKD-------LEGKVVVITGSSTGLGKAMAIRFATEKAK-VVVNYRSKEEEANSVLEEI
>HBDH           M--D-------FSGKNVWVTGAGKGIGYATALAFVEAGAK-VTGFDQAFTQEQYPFATEV
>PGDH           M--H-------VNGKVALVTGAAQGIGRAFAEALLLKGAK-VALVDWNLEAGVQCKAALD
                *           .. .  ...  ..*          .  .      .
                       60        70        80        90              100
>DlADH          NPKVNITFHTYDVTVPVAES-KKLLKKIFDQLKTVDILINGAGILDD--------HQIER
>DmADH          NPKVTVTFYPYDVTVPIAET-TKLLKTIFAQLKTVDVLINGAGILDD--------HQIER
>HSD            RELG-DAARYQHLDVTIEEDWQRVVAYAREEFGSVDGLVNNAGISTGMFLETESVERFRK
>RDH            AELGQN-AFALQVDLMQADQVDNLLQGILQLTGRLDIFHANAGAYIGGPVAEGDPDVWDR
>SDH            AEYGESMAYGFGADATSEQSCLALSRGVDEIFGRVDLLVYSAGIAKAAFISDFQLGDFDR
>GLU DH         KKVG-GEAIAVKGDVTVESDVINLVQSSIKEFGKLDVMINNAGMENPVSSHEMSLSDWNK
>HBDH           MDVA-DAAQVAQVC-------QRLLA  ---ETERLDALVNAAGILRMGATDQLSKEDWQQ
>PGDH           EQFEPQKTLFIQCDVADQQQLRDTFRKVVDHFGRLDILVNNAGVNNE--------KNWEK
                                              .* .  .**                  .
                     110       120       130       140       150       160
>DlADH          TIAINFTGLVNTTTAILDFWDKRKGGPGGIIANICSVTGFNAIHQVPVYSASKAAVVSFT
>DmADH          TIAVNYTGLVNTTTAILDFWDKRKGGPGGIICNIGSVTGFNAIYQVPVYSGTKAAVVNFT
>HSD            VVDINLTGVFIGMKTVIPAM--KDAG-GGSIVNISSAAGLMGLALTSSYGASKWGVRGLS
>RDH            VLHLNINAAFRCVRSVLPHLLAQKS---GDIIFTAVIAGVVIW--EPVYTASKFAVQAFV
>SDH            SLQVNLVGYFLCAREFSRLMIR--DGIQGRIIQINSKSGKVGSKHNSGYSAAKFGGVGLT
>GLU DH         VIDTNLTGAFLGSREAIKYF--VENDIKGTVINMSSVHEKIPWPLFVHYAASKGGMKLMT
>HBDH           TFAVNVGGAFNLFQQTMNQF--RRQR-GGAIVTVASDAAHTPRIGMSAYGASKAALKSLA
>PGDH           TLQINLVSVISGTYLGLDYMSKQNGGEGGIIINMSSLAGLMPVAQQPVYCASKHGIVGFT
                    *  .                      * .       .         * ..* .   .
                     170       180       190       200       210
>DlADH          NSLAKLAPIT--GVTAYSINPG-ITRTPLVHTF----NSWLDVEPRVAELLLSHPTQTSE
>DmADH          SSLAKLAPIT--GVTAYTVNPG-ITRTTLVHTF----NSWLDVEPQVAEKLLAHPTQSSL
>HSD            KLAAV--ELGTDRIRVNSVHPG-MTYTPMTAE--TGIRQGEGNYPN--------TPMGRV
>RDH            HTTRR--QVAQYGVRVGAVLPG-----PVVTALLDDWPK-----AKMDEALADGSLM---
>SDH            QSLAL--DLAEYGITVHSLMLGNLLKSPMFQSLLPQYATKLGIKPDQVEQYYIDKVPLKR
>GLU DH         ETLAL--EYAPKGIRVNNIGPG-AINTPINAEKFADPEQRADVESM--------IPMGYI
>HBDH           LSVGL--ELAGSGVRCNVVSPG-STDTDMQRTLWVSDDAEEQRIRGFGEQFKLGIPLGKI
>PGDH           RSAALAANLMNSGVRLNAICPG-FVNTAILES--IEKEENMGQYIEYKDHIKDMIKYCGI
                          .    .  *      .  .
                       220       230       240       250
>DlADH          QCGQNFVKA----IEANKNGAIWKLDLGTLEAIEWT-----KHWDSHI
>DmADH          ACAENFVKA----IELNQNGAIWKLDLGTLEAIQWT-----KHWDSGI
>HSD            GNEPGEIAGAVVKLLSDTSSYVTGAELAVDGG--WTTGPTVKY-VMGQ
>RDH            --QPIEVAESVLFMVTRSKNVTVRDIVILPNSVDL-------------
>SDH            GCDYQDVLNMLLFYASPKASYCTGQSINVTGG-------QVMF-----
>GLU DH         G-EPEEIAAVAAWLASSEASYVTGITLFADGG--MTQYPSFQA-GRG-
>HBDH           AR-PQEIANTILFLASDLASHITLQDIVVDGG--ST---------LGA
>PGDH           L-DPPLIANGLITLIEDDALNGAIMKITTSKGIHFQDYDTTPFQAKTQ
                      . .                 .    .

Key: . = conserved substitution
     * = identity
```

Figure 1–1: Alignment of some of the short chain dehydrogenases, showing the conserved residues. For a more complete study see Persson *et al*, 1991.

Enzymes	Species	Abbreviations	Reference
Alcohol dehydrogenase	*Drosophila lebanonensis*	DlADH	Villarroya *et al.*, 1989
Alcohol dehydrogenase	*Drosophila melanogaster*	DmADH	Thatcher, 1980
3α,20β-hydroxysteroid dehydrogenase	*Streptomyces hydrogenans*	HSD	Marekov *et al.*, 1990
Ribitol dehydrogenase	*Enterobacter aerogenes*	RDH	Irie *et al.*, 1987
Sorbitol-6-phosphate dehydrogenase	*E.coli*	SDH	Yamada and Saier, 1987
Glucose dehydrogenase	*Bacillus megaterium*	GLU DH	Jany *et al.*, 1984
2,3-Dihydro-2,3-dihydroxybenzoate dehydrogenase	*E.coli*	HBDH	Lui *et al.*, 1989
15-hydroxyprostaglandin dehydrogenase	Human	PGDH	Krook *et al.*, 1990

Table 1–1: List of short chain dehydrogenases aligned overleaf.

has been speculated that this may be a result of NAD^+ binding and that the apo-structure would reveal two domains (Krook *et al.*, 1992). Each domain has a β sheet comprising seven parallel strands with three parallel helices on each side. This is the classic dinucleotide binding domain or Rossmann fold and is composed of approximately 140 residues. The carboxy-terminal segment of the HSD molecule consists of 67 residues which form a helix, αG, which is packed closely and parallel to αB, a large loop λG (which has weak density), and a final short β strand, βG. The 16 carboxy-terminal residues were badly defined in the electron density map. The molecule is a tetramer (see Figure 1–2) and the monomer-monomer interactions are as follows:

- **P-axis** packing of αG. λG and carboxyl-terminal loop and two amino termini associate about this axis.

- **R-axis** antiparallel association between the carboxy-terminal region and 3 residues in the loop λE. The R-axis forms the edge of the cleft where the steroid is thought to bind, therefore explaining why a tetrameric form of this enzyme is essential for catalysis (Carrea *et al.*, 1984).

- **Q-axis** βD-λD-αE, αE, βE-λE-αF and αF are packed closely together across this axis. This monomer-monomer interface holds the cofactors in the closest proximity.

The crystal structure of HSD shows an unusual method of binding of the NAD(H) which has not been observed in other dehydrogenases (Ghosh *et al.*, 1991). The NAD(H) in HSD is positioned nearer the exterior of the protein than positioning observed for medium chain dehydrogenases (see Figure 1–3). This

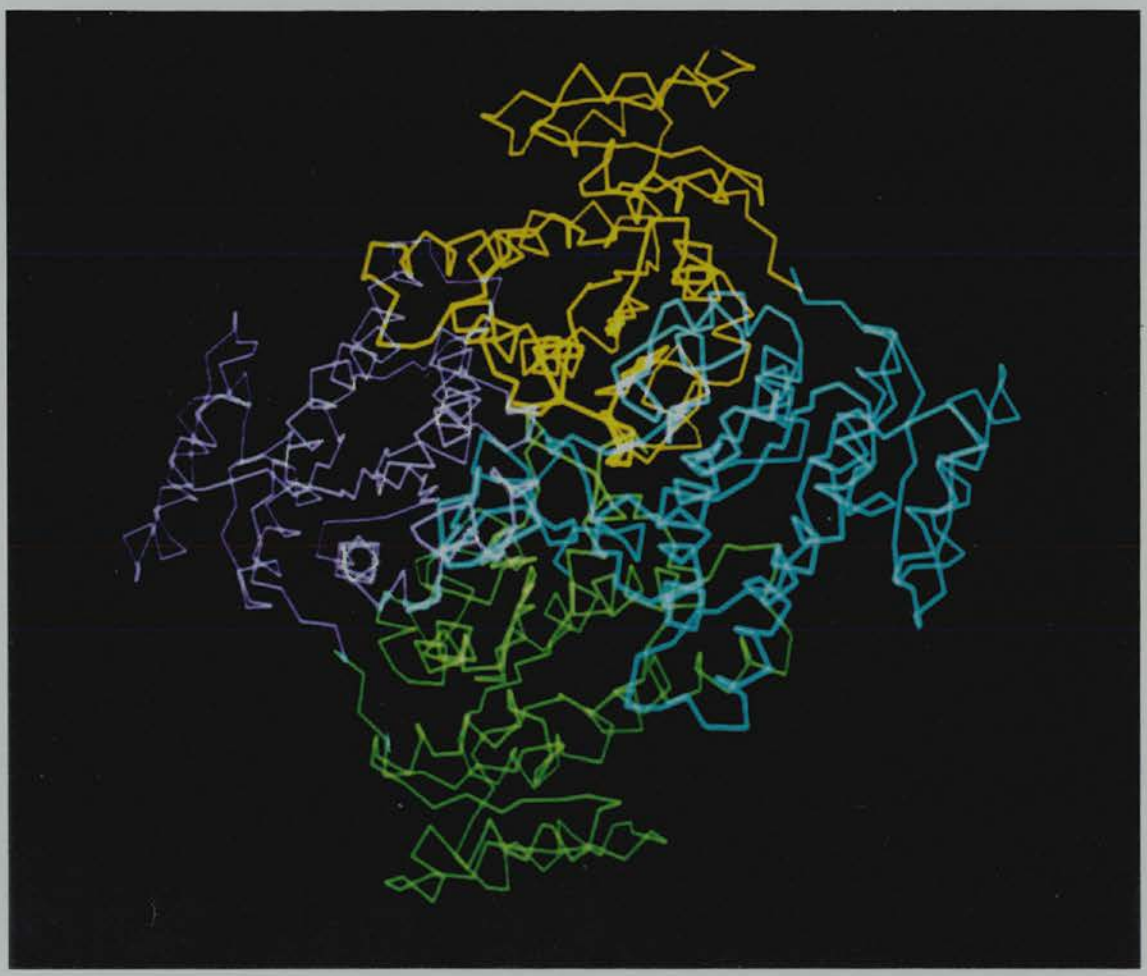
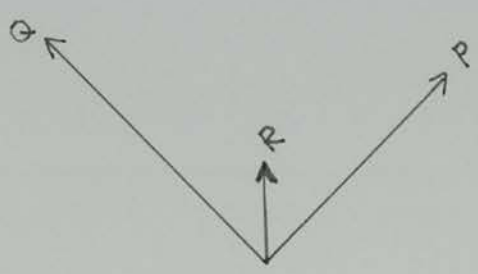


Figure 1–2: Plate showing the HSD tetramer



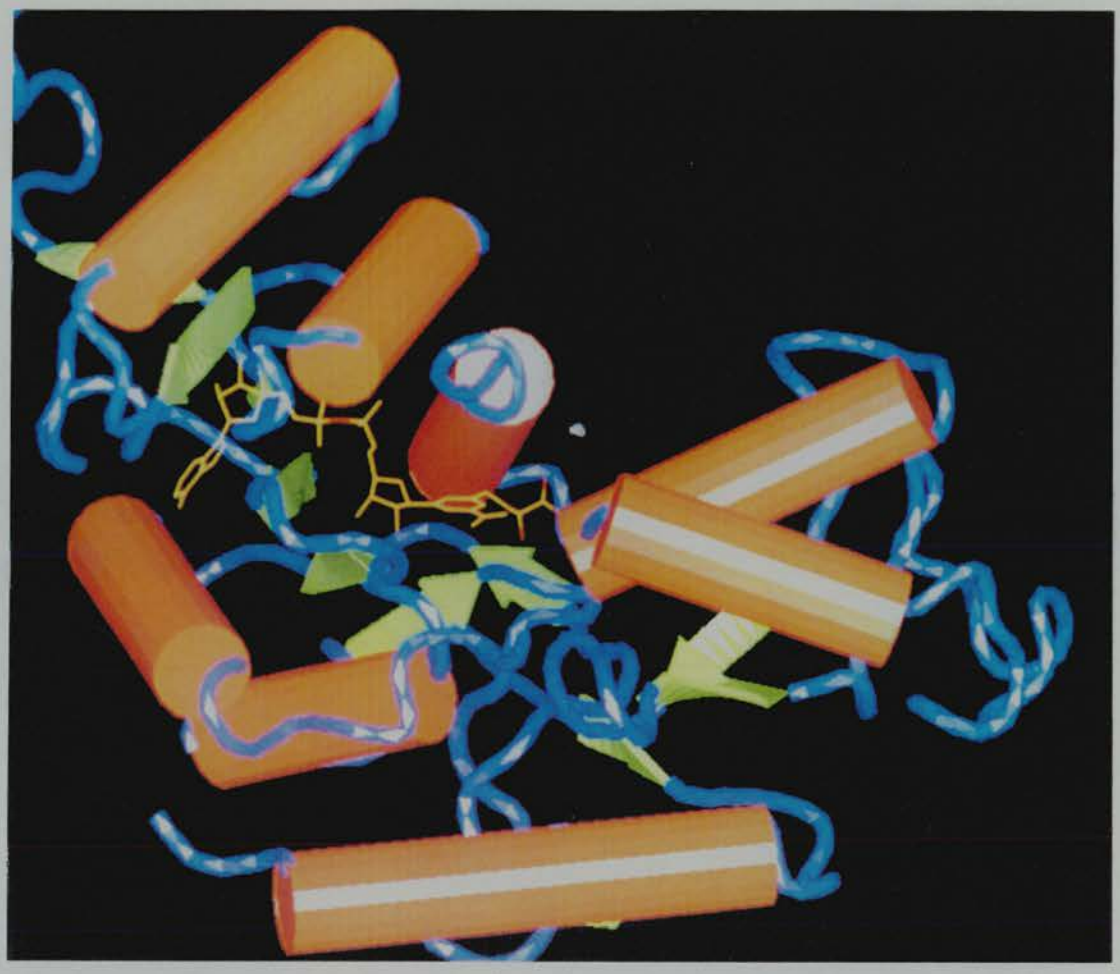


Figure 1–3: Picture of NAD^+ binding to LDH, this is typical of the binding found in medium chain dehydrogenases. The NAD^+ lies along one edge of the β sheet which forms the Rossmann fold. The adenosine part of the NAD^+ lies close to the first β strand in the fold (see also Figure 1–4)in HSD the nicotinamide ring lies in this position.

result might be an artifact of crystal packing or a real difference in cofactor binding between medium and short chain dehydrogenases. Confirmation of this result awaits further short chain dehydrogenase structures.

The conserved regions of sequence within the short chain dehydrogenases are analysed with respect to the HSD structure. Residues 5-15 cover the glycine rich turn, sequence motif GXGXXG or GXGXXA, which is found in many dinucleotide binding domains (Wierenga *et al.*, 1985). This tight turn allows the central region of the NAD^+ to lie close to the protein framework and hydrogen bond,

directly or indirectly, to the adenosine ribose. The Rossmann fold can bind either NAD^+ or $NADP^+$ as cofactor (Baker *et al.*, 1992).

A conserved Asp at position 38 in *Drosophila* (see figure 1–1) has been shown to be important in cofactor specificity (Chen *et al*, 1991). However recent studies on $NADP^+$ and dual specific glutamate dehydrogenases (GDH) have identified an equivalent acidic residue. In HSD, Asp-38 is within hydrogen bonding distance of the carboxamide group of the NAD^+. In other dehydrogenases this residue has been proposed to bind to the 2'-OH adenosine ribose (Brändén and Tooze, 1991). If the NAD^+ in the HSD structure was moved further into the interior of the enzyme this negatively charged residue would come close to the pyrophosphate moiety of the NAD^+ where it may be involved in recognition of the 2' phosphate, either directly or indirectly.

There are three strictly conserved non-glycine residues in the short chain dehydrogenases, Asp-63, Tyr-151 and Lys-155 (Persson *et al.*, 1991). The strict conservation of Asp-63 is not obvious from our alignment (see Figure 1–1) and decreasing gap weights did not improve the alignment significantly. Ghosh *et al.*, (1991) do not propose a function for this residue in the HSD structure (Asp-63), indeed the residue is found near the surface of the molecule in a stretch of β strand. Why is Asp-63 strictly conserved in all short chain dehydrogenases ? If the positioning of Asp-65 in HSD is common to all dehydrogenases then it is possible this residue is important in orientating and maintaining packing of βC against αC, and hence maintaining the positions of Asp-37 and Glu-87 with respect to the coenzyme. Would other negatively charged residues be acceptable ? The strongly conserved Asp-63, indicates an interaction with substrate or cofactor, since it is rare that a charged side chain is conserved to maintain structure. However, this does not agree with the HSD structure positioning of this Asp residue suggesting that either the Asp has been identified incorrectly as a highly conserved residue or that the HSD structure has wrongly positioned this residue. An alternative explanation for this region of the short chain structures is that a charged residue is needed to bind the cofactor, but that this residue can be either an Asp or a Glu. The HSD structure has a Glu in the loop region

between, βC and αD, Glu-65 which may hydrogen bond to the cofactor. If a Glu or another charged residue is present, in all short chain dehydrogenases it suggests a common system for cofactor binding in these enzymes. However, there is no charged residue in this position in the DADH sequences.

A Tyr has also been found to be conserved in all short chain dehydrogenases (Krook *et al.*, 1990; Persson *et al.*, 1991). This tyrosine has been shown to be essential for the activity of HSD (Krook *et al.*, 1992), PGDH (Ensor and Tai, 1991) and DmADH (Chen *et al.*, 1992).

Both Tyr 151 and Lys 155 (DmADH nomenclature), lie in the central region of the short chain dehydrogenases which has long been highlighted as a highly conserved region (Jörnvall *et al.*, 1981; Villarroya *et al.*, 1989; Persson *et al*, 1991). This region forms the putative active site of HSD (Ghosh *et al.*, 1991). It lies in a cleft between the two subunits, forming ligands with both domains and from different subunits. The HSD structure shows Ser 139 also lies in the putative active site pocket, this residue is conserved in nearly all of the short chain dehydrogenases. Together, Ser 139, Tyr 152 and Lys 156 (HSD nomenclature) form a polar pocket on the surface of the HSD molecule. It has been proposed that the Tyr 152 and Ser 139 in HSD hydrogen bond to the steroid substrate. No role is proposed for the conserved Lys residue.

1.3 Dinucleotide binding domains

Proteins that bind NAD^+ and $NADP^+$ have been found to have a very similar three dimensional fold in spite of having low sequence identities. This fold is the dinucleotide binding domain or Rossmann fold, and is constructed from 6 parallel β strands with α helices on each side of the sheet. This $\alpha\beta$ structure is built from two $\beta\alpha\beta\alpha\beta$ motifs or mononucleotide binding domains. Comparisons of a number of structures of the medium chain dehydrogenases reveal a core structure, which is defined as the complete β1-αA-β2 motif (see Figure 1–4), the major parts of αB and αD and the remaining four β strands (see Brändèn and

Figure 1–4: The β1-αA-β2 motif of the mononucleotide binding domain. The position of the NAD^+ is typical of binding observed in medium chain dehydrogenases.

Tooze, 1992). This core structure has an overall r.m.s. deviation of 2Å (between LDH, LADH and GAPDH). In most dehydrogenases the NAD^+ and $NADP^+$ bind in a similar conformation to the Rossmann fold (see Figure 1–3) the exceptions are GAPDH (Skarzyuski *et al.*, 1987) and HSD (Ghosh *et al.*, 1991). Generally, the NAD^+ binds in the cleft formed by the carboxyl ends of the β strands 1 and 4, with the pyrophosphate group straddling the β sheet. The adenosine binds to the first mononucleotide binding domain and the ribose of the nicotinamide part of the NAD^+ to the second domain. The transfer of the hydrogen to the C4 position on the nicotinamide ring is stereospecific. Due to the carboxyamide substitution at C3, the two hydrogen positions at C4 are not equivalent. Therefore when the dehydrogenase transfers a hydrogen to the position above the ring the dehydrogenase is termed class A, and when the

hydrogen transfer is to the position below the ring it is a class B dehydrogenase. The NAD^+ is positioned close to the substrate and it is the interactions of the carboxyamide group of the NAD^+ with the protein that determines the stereospecificity of the hydrogen transfer.

Although the sequence identity for the dinucleotide binding domain is low, several conserved residues have been highlighted and a fingerprint constructed such that the $\beta\alpha\beta$ structure can be predicted from sequence information alone (Wierenga *et al.*, 1985). A GXGXXG sequence motif lies in the turn at the end of the first β strand and before the α helix. The first Gly of this motif is essential in conserving the tight turn between the β strand and the α helix (in the *Drosophila lebanonensis* this Gly is changed to an Ala). The second Gly allows for the close interaction of the dinucleotide to the protein and the third allows for the close interaction of the two β strands and the α helix. The helix dipole also contributes to coenzyme binding by interacting favourably with the pyrophosphate moiety. Other conserved features of the fingerprint region (Wierenga *et al.*, 1985) are :

- A hydrophilic residue at N-terminus of the first β strand. The function of this residue is not known.

- 6 small residues in the hydrophobic core.

- A negatively charged residue at the C-terminus of the second β strand. This residue forms a hydrogen bond to the 2'-hydroxyl group of adenine ribose group on NAD^+.

A small number of enzymes utilize $NADP^+$ as a cofactor (which has an additional phosphate group esterified to the 2'-hydroxyl group to the AMP part of the NAD^+). A similar fingerprint region can be identified for these enzymes, with significant differences:

- The negatively charged residue at the C-terminal end of the second β strand is replaced, presumably to accommodate 2'phosphate group.

- The third Gly in the conserved motif is replaced by an Ala.
- There is almost always another Ala in the 4 residues following the α helix (towards the C-terminus).

The mutations that are needed to change the cofactor specificity of an enzyme from $NADP^+$ to NAD^+ (Scrutton *et al.*, 1990) are all found in the $\beta\alpha\beta$ fold of the above fingerprint region.

1.4 *Drosophila* ADH

The alcohol dehydrogenase enzymes from *Drosophila* are dimeric proteins with between 253-255 residues in each polypeptide chain and with no bound metal ions.

ADH is produced in high levels in both *Drosophila* larvae and in adult flies. The pattern of expression of *Adh* is tissue specific and expression is different in different *Drosophila* species (Ursprung *et al.*, 1970; Dunn *et al.*, 1969; Knowles and Fristrom, 1967; Alberola *et al.*, 1987). Most work has been carried out on the *Drosophila melanogaster* species, a species which has a high alcohol dehydrogenase activity. In fact, *Drosophila melanogaster* thrive in areas of high alcohol content. DADH can also catalyse the next reaction, that is they show aldehyde dehydrogenase activity (Heinstra *et al.*, 1983).

1.4.1 Allelic variants

Natural populations of DADH show many allelic variants. *Adh*-null alleles occur very rarely in natural populations and it has been found that exposure to even low concentrations of ethanol can destroy these *Drosophila*. However, in certain environments, unsaturated alcohols are converted to highly toxic ketones which kill the ADH-positive individuals. ADH-null individuals can survive in these conditions (van Deldon, 1982)

Alleloenzyme	Position	ADH-S	Change
ADH-F'	51	Ala	Glu
ADH-F	192	Lys	Thr
ADH-UF	8	Asn	Ala
	45	Ala	Asp
	192	Lys	Thr
ADH-D	192	Lys	Thr
	232	Gly	Glu
ADH-FChD	192	Lys	Thr
	214	Pro	Ser

Table 1–2: Table showing the alleloenzymes of *Drosophila melanogaster* that have been identified and characterized

Individual flies produce isozymes: ADH-1, ADH-2 and ADH-3. Studies show that these forms of the enzyme are due to the covalent attachment of NAD^+-carbonyl adducts to the subunits (Winberg, 1988).

The *Adh* locus in *Drosophila melanogaster* is polymorphic and several isozymes and alleloenzymes have been identified. These are classified according to their rate of mobility towards the anode on an electrophoresis gel. The alcohol dehydrogenase from different species as well as the alleloenzymic variants from *Drosophila melanogaster* have been sequenced and biochemically characterized (Batterham *et al.*, 1983; Thatcher, 1977, 1980; Chambers *et al.*, 1981a, b, 1984; Winberg *et al.*, 1982a, b, 1983, 1985, 1986; Juan and Gonzalez-Duarte, 1980, 1981; Chambers *et al.*, 1984; Gibson *et al.*, 1980; Vilageliu and Gonzalez-Duarte, 1984; Atrian and Gonzalez-Duarte, 1985; Villarroya, *et al.*, 1989; Winberg and McKinley-McKee, 1988a, b, 1992)

Table 1–2 gives a list of the reported alleloenzymes for *Drosophila melanogaster.* The most common alleloenzymes are the ADH-F and ADH-S enzymes distinguishable by a single amino acid change, Lys-192 in ADH-S is changed to Thr-192 in ADH-F (Retzio and Thatcher, 1979; Thatcher, 1980). This abundance of natural variants gives us information on the effect of a single residue change on the enzyme's activity and/or stability. It also indicates important segments of the molecule, strictly conserved residues and suggests positions suitable for site-directed mutagenesis experiments.

Species	Abbreviations	Reference
D. melanogaster-FChD	D.mel	Chambers, 1984
D. lebanonensis	D.leb	Albalat and Gonzalez-Duarte, 1989
D. mulleri	D.mul	Fischer and Maniatis, 1985
D. mojavensis	D.moja	Atkinson *et al.*, 1988
D. navojoa	D.navo	Weaver *et al.*, 1989
D. simulans	D.sim	Cohn *et al.*, 1984
D. sechellia	D.seche	Coyne and Kreitman, 1986
D. mauritiana	D.maur	Cohn *et al.*, 1984
D. orena	D.oren	Bodmer and Ashburner, 1984

Table 1–3: List of *Drosophila* ADH sequences included in alignment

1.4.2 Species variations

There have now been a number of DADH enzymes sequenced both at the protein and cDNA level (Thatcher, 1980; Villarroya *et al.*, 1989; Benyajati *et al.*, 1981; Schaeffer and Aquadro, 1987; Rowan and Dickerson, 1988; Fisher and Maniatis, 1985; Atkinson *et al.*, 1988). Sequence comparisons show that the DADH enzymes exhibit a number of differences, but share an identity of 62.5% between sequences or 80.6% if conservative substitutions are allowed (Winberg and McKinley-McKee, 1992). These identities are spread throughout the total protein sequence with glycine being the most conserved residue in the *Drosophila* enzymes (Jörnvall *et al.*, 1984b). Only the frequency of two amino acids, cysteine and tryptophan, have been maintained. An alignment of all *Drosophila* ADH sequences is shown in figure 1–5.

1.4.3 *Drosophila lebanonensis*

ADH from *Drosophila lebanonensis* shows a high tolerance to alcohol and a better ability to utilize ethanol as a food resource than that observed for the *D. melanogaster* enzyme (David *et al.*, 1979). Phylogenetically *D. lebanonensis* is distantly related to *D. melanogaster* (Vilageliu and Gonzalez-Duarte, 1984). The alcohol dehydrogenase from *D. lebanonensis* has been sequenced (Villarroya *et al.*, 1989). Each subunit has 254 residues in the polypeptide chain, two residues less than the *D.melanogaster* Chateau Douglas alleloenzyme. Like the DmADH,

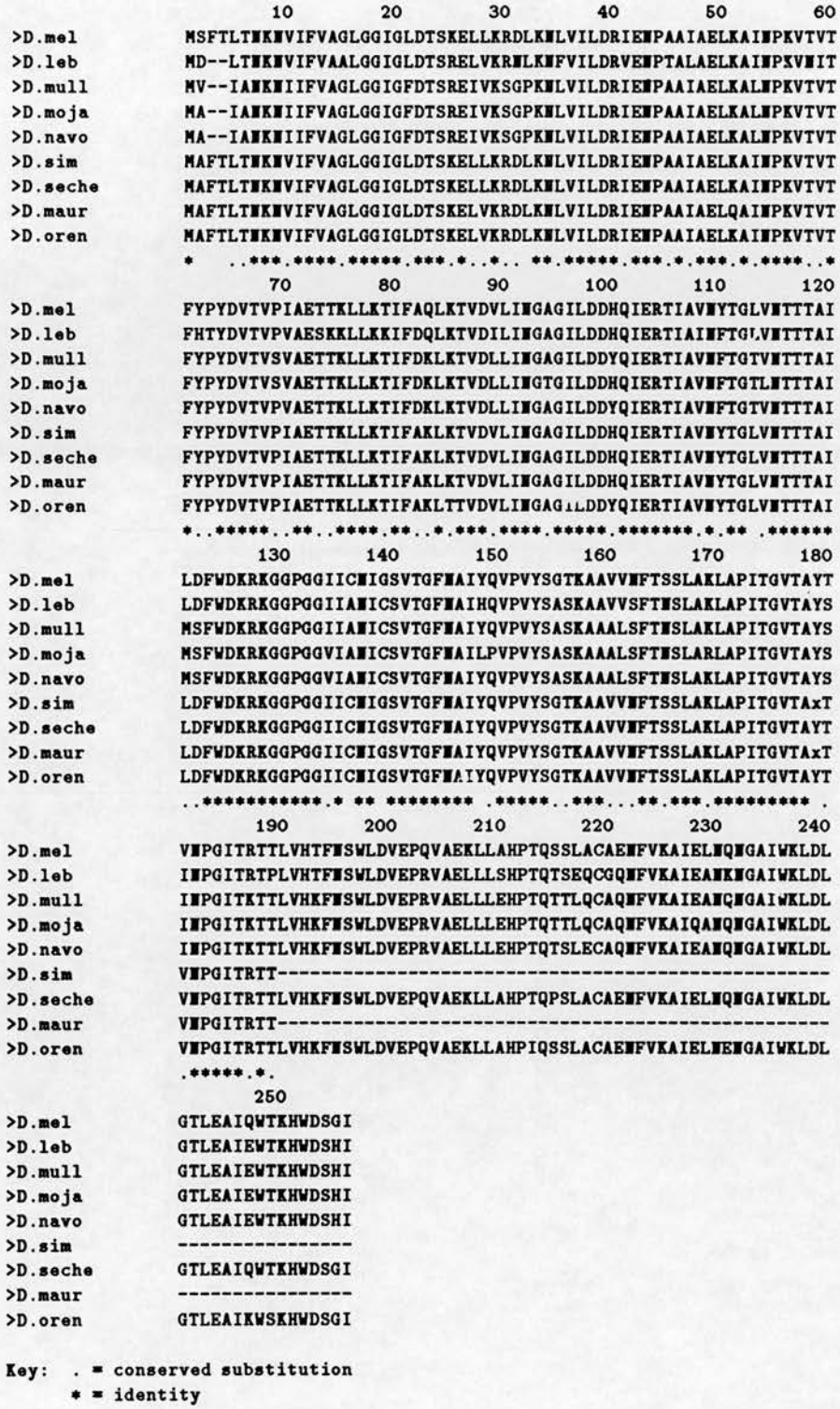


Figure 1–5: Sequence alignment of some sequences of *Drosophila* alcohol dehydrogenase. The *D. melanogaster* sequence is that of the Fast Chateau Douglas alloenzyme. The sequences for *D. simulans* and *D. mauritiana* are incomplete.

```
                10        20        30        40        50        60
>D.mel          MSFTLTNKNVIFVAGLGGIGLDTSKELLKRDLKNLVILDRIENPAAIAELKAINPKVTVT
>D.leb          MD--LTNKNVIFVAALGGIGLDTSRELVKRNLKNFVILDRVENPTALAELKAINPKVNIT
                *.  **********.*********.**.**.***.*****.***.*.**********..*
                      70        80        90        100       110       120
>D.mel          FYPYDVTVPIAETTKLLKTIFAQLKTVDVLINGAGILDDHQIERTIAVNYTGLVNTTTAI
>D.leb          FHTYDVTVPVAESKKLLKKIFDQLKTVDILINGAGILDDHQIERTIAINFTGLVNTTTAI
                *..******.**..****.**.******.******************.*.**********
                      130       140       150       160       170       180
>D.mel          LDFWDKRKGGPGGIICNIGSVTGFNAIYQVPVYSGTKAAVVNFTSSLAKLAPITGVTAYT
>D.leb          LDFWDKRKGGPGGIIANICSVTGFNAIHQVPVYSASKAAVVSFTNSLAKLAPITGVTAYS
                ************** ** ********.******..*****.**.**************.
                      190       200       210       220       230       240
>D.mel          VNPGITRTTLVHTFNSWLDVEPQVAEKLLAHPTQSSLACAENFVKAIELNQNGAIWKLDL
>D.leb          INPGITRTPLVHTFNSWLDVEPRVAELLLSHPTQTSEQCGQNFVKAIEANKNGAIWKLDL
                .*******.*************.*** **.****.* .*..******* *.*********
                      250
>D.mel          GTLEAIQWTKHWDSGI
>D.leb          GTLEAIEWTKHWDSHI
                ******.******* *

Key:   . = conserved substitutions
       * = identity
```

Figure 1–6: Pairwise alignment of the ADH-FChD alleloenzyme of *D. melanogaster* ADH with the *D. lebanonensis* ADH

DlADH has an acetylated N-terminal methionine. DlADH differs from DmADH-FChD at 43 positions, these changes are spread along the entire length of the sequences (see Figure 1–6).

Comparisons of the *D. lebanonensis* sequence with ribitol dehydrogenase confirm the similarity which had been identified with the *D. melanogaster* (Jörnvall *et al.*, 1981). Sequence comparisons with other *Drosophila* ADH's suggest that they all share a similar tertiary structure, although the extent of any similarity is not known. For example, c.d. experiments indicate that *D. lebanonensis* has 10% more β strand structure than *D. melanogaster* (Ribas de Pouplana *et al.*, 1991).

Some of the biochemical properties of the alcohol dehydrogenase from DlADH have been determined (Winberg *et al.*, 1986) and compared to the biochemical data from DmADH. The *D. lebanonensis* alcohol dehydrogenase has three active zones on an electrophoresis gel. The isoelectric points of these zones are 7.95,

7.50 and 7.15 which are nearer to the cathode than the DmADH^S-5 isozyme. As with *D. melanogaster*, the ADH-5 form of *D. lebanonensis* ADH, can be converted to ADH-1 and ADH-3 forms by reacting with NAD^+ and a ketone (Winberg and McKinley-McKee, 1988a).

It has been shown that the substrate specificity of DlADH is similar to that of other *Drosophila* species (Chambers *et al.*, 1981; Juan and Gonzalez-Duarte, 1981; Oakeshott *et al.*, 1982; Winberg *et al.*, 1982; Hovik *et al.*, 1984) with secondary alcohols being better substrates than primary alcohols.

1.4.4 Comparisons

The effect of different mutations observed in some DADH species (or alleloenzymes), on the enzyme activity have been studied and these indicate important residues involved in catalysis.

In the cofactor binding region of the DlADH, the first Gly residue of the GXGXXG motif characteristic of dinucleotide binding folds, is mutated to an Ala. It has been shown that site-directed mutagenesis of the DmADH Gly 14 to Ala 14, gives a slightly reduced enzymatic activity (Chen *et al.*, 1990). The substitution of larger and charged residues at this position prevent NAD^+ binding and render the enzyme inactive.

ADH-S and ADH-F alleloenzymes vary in many, if not all of their biochemical characteristics. ADH-S binds NAD^+ strongly and since the rate limiting step of the oxidation of the secondary alcohols by DADH is the dissociation of the ternary complex, the reaction of ADH-S with secondary alcohols is slow. In contrast, ADH-F binds NAD^+ weakly due to the additional charge at position 192 and therefore the reaction of ADH-F with secondary alcohols is quick. The implication is that residue 192 is either directly or indirectly involved in coenzyme binding and this is supported by results from inhibitor studies (Winberg *et al.*, 1982; Hovik *et al.*, 1984). The ADH-F enzyme has a reaction rate with ethanol that is twice that of ADH-S which may be due to the quicker dissociation rate of the enzyme-NADH complex; because residue 192 affects

hydride transfer or possibly some other factor. It has also been shown that the ADH-S variant is more thermodynamically stable than the ADH-F variant (Vigue and Johnson, 1973; Day *et al.*, 1974; Thatcher and Sheik, 1981).

ADH-FChD (the same as ADH-71K alleloenzyme) is a heat stable variant which has an electrophoretic mobility identical to the ADH-F alleloenzyme (Thörig *et al.*, 1975). ADH-FChD differs from ADH-F in one amino acid substitution, residue 214 which is a Pro in all DADH except for ADH-FChD and the ADH from *D.orena* where residue 214 is a Ser. ADH-FChD has a slower reaction rate, similar to that for ADH-S, even though it contains a Thr-192 (Heinstra *et al.*, 1988). This implies that residue 214 is involved in cofactor binding either directly or indirectly. Also, that this residue increases the thermal stability of the enzyme in some way.

D.simulans differs from ADH-S at positions 1 and 82, it shows much slower reaction rates. Since residue 1 is not involved in catalysis or cofactor binding (Winberg and McKinley-McKee, 1992), the slower dissociation rate of the enzyme-NADH complex for *D.simulans* (Juan and Gonzalez-Duarte, 1981) must be a result of the change Gln82Lys. Residue 82 is located in the N-terminal part of β strand 4 in the dinucleotide binding domain.

Substrate specificity for many of these variants (Winberg *et al.*, 1982b; Hovik *et al.*, 1984) show that the DADH's are more active with secondary alcohol substrates, hence a topology for the active site should contain a number of hydrophobic residues. Substrate specificities have been studied by calculating V_{max}/K_M (where K_M is the apparent dissociation constant that can be treated as the overall dissociation constant of all enzyme-bound species and where V_{max} is the maximum velocity) for various substrates and inhibitors. This function reflects the reaction rate at low substrate concentrations. Its advantages are that it is unaffected by non-productive binding or by accumulation of intermediates (Fersht, 1977). It was found that V_{max}/K_M for ethanol was almost the same for all DADH enzymes tested and it was 20 times higher for propan-2-ol than for ethanol, indicating the DADH's preference for secondary alcohols. The difference in binding between a primary and secondary alcohol is about -7kJ/mol. which is

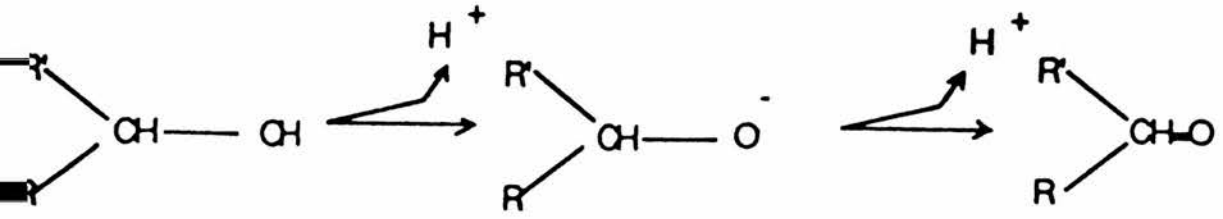


Figure 1–7: Reaction mechanisms for oxidation of primary and secondary alcohols

the energy range for an additional hydrophobic interaction, which indicates that both methyl groups of the secondary alcohol interact hydrophobically with the DADH enzyme. As expected, the activity of the DADH decreases for charged and polar alcohols (Winberg *et al.*, 1982a, 1986; Hovik *et al.*, 1984).

1.4.5 Catalysis

DADH catalyses the oxidation of aliphatic primary and secondary alcohols to aldehydes and ketones with the concomitant reduction of NAD^+ to NADH. These enzymes show a preference for secondary alcohols.

Studies have indicated that the reaction with secondary alcohols is an ordered reaction pathway with the rate limiting step being the dissociation of the enzyme-NADH complex (Winberg *et al.*, 1982; Hovik *et al.*, 1984). For primary alcohols a similar mechanism was proposed, but with the rate limiting step being the formation of the ternary complex (Winberg, *et al.*, 1986). However, a rapid equilibrium random (RER) mechanism has now been proposed for the oxidation of primary alcohol (Chambers, 1991), since it has been observed that NADH acts as a competitive inhibitor with respect to ethanol (Heinstra, *et al.*, 1988). The RER mechanism implies no obligatory order and the chemical steps are slower than the binding of reagents.

In the mammalian liver alcohol dehydrogenases, an ordered mechanism is

necessary because the ethanol binds to the enzyme as an alcoholate ion. The proton liberated as a result of this ion formation is removed *via* a charge relay system to the protein. Therefore, the NAD^+ must bind before the substrate. Mammalian alcohol dehydrogenases (LADH) have an active site zinc atom which plays a crucial role in this reaction mechanism. The zinc atom affects ionization properties of the enzyme bound alcohol, by aiding proton transfer to solution, and hydride ion transfer to NAD^+. *Drosophila* alcohol dehydrogenase has no active site zinc atom (Place *et al.*, 1980; Moxon *et al.*, 1985; Winberg *et al.*, 1986), but kinetic studies of the pH dependence of the *Drosophila* alcohol dehydrogenase (Winberg and McKinley-McKee, 1988b) observe some similarities with the pH dependence for the liver alcohol dehydrogenase enzyme (Theorell and McKinley-McKee, 1961; Daziel, 1963; Kvassman and Pettersson, 1980; Pettersson, 1986).

Winberg *et al.*, (1991) have proposed a possible mechanism for the reaction mechanism of alcohol dehydrogenase from *Drosophila* and shown that it is significantly different to that of the horse liver alcohol dehydrogenase mechanism.

The horse liver alcohol dehydrogenase decreases the pK_a for the enzyme-bound alcohol substrate so that at the physiological pH, the substrate is present predominantly as alcoholate ions in the productive ternary complexes. A negative charge on the substrate facilitates hydride transfer to the NAD^+.

The *Drosophila* reaction is an ordered ternary complex mechanism. The ternary complex formation with an alcohol substrate is tightened by ionization of an enzymatic group with a pK_a close to 7.6 (Kvassman and Pettersson, 1980; Winberg and McKinley-McKee, 1988b). The binding of aldehyde is not affected by a change in pH (Winberg and McKinley-McKee, 1988b). This ionizing group (which has a pK_a 7.6), possibly acts as a nucleophilic catalyst of the hydride transfer, by increasing the negative charge on the substrate, when it is in the ternary complex. The alcohol binds to the unprotonated form of the ionizing group in the enzyme (this is equivalent to the alcoholate ion binding to the protonated group). The percentage of alcoholate ion formed would be small unless the substrate binding group carries a negative charge in its unprotonated

state, for example, a cysteinyl or tyrosyl residue could act as the substrate binding group. Evidence from site-directed mutagenesis studies indicates that the cysteine residues in DADH are not catalytic (Chen *et al.*, 1990) but that the Tyr 152 is involved in catalysis (Chen *et al.*, 1992).

The suggestion that a Tyr is involved in substrate binding and as a nucleophilic catalyst is consistent with the Tyr 152, in DmADH, and 151, in DlADH, being conserved in all *Drosophila* species so far examined (Villarroya *et al.*, 1989) and this residue being highlighted as a conserved residue in all short chain dehydrogenases (Krook *et al.*, 1990; Marekov *et al.*, 1990; Ensor and Tai, 1991). Site directed mutagenesis studies have shown that Tyr 152 is necessary for activity in HSD (Krook *et al.*, 1992), PGDH (Ensor and Tai, 1991) and DADH (Chen *et al.*, 1992).

On the basis of this proposed mechanism (Winberg and McKinley-McKee, 1992) a topology has been suggested for the active site of *Drosophila* alcohol dehydrogenase (see Figure 1–8). The observation of the reaction mechanism for primary alcohols for the *Drosophila* enzyme implies that the mechanisms employed for the medium chain, zinc containing enzymes and the short chain enzymes are different.

Ribas de Pouplana and Fothergill-Gilmore (1993) have recently proposed a mechanism for the action of DADH (see Figure 1–9), on the basis of an homology modeling experiment; a mechanism reminiscent of that adopted by LDH (Adams, 1987). They propose that the NAD^+ binds to the enzyme and this causes a conformational change which allows the substrate to bind to Lys 156. The C-terminal loop then folds down to cover the active site and brings His 250 (which lies in this loop) close to the active site. His 250 acts as a base which promotes the hydride transfer to the NAD^+ (the imidazolium ion is stabilized by Asp 97). The loop then opens and a proton is released into the solvent. This mechanism fulfills a point made by Winberg and McKinley-McKee; that the alcohol binds to an unprotonated form of the enzyme group, which has a pK_a=7.6 (Winberg and McKinley-McKee, 1988b). The observation that proteolysis of a C-terminal portion of the DmADH enzyme, renders the enzyme

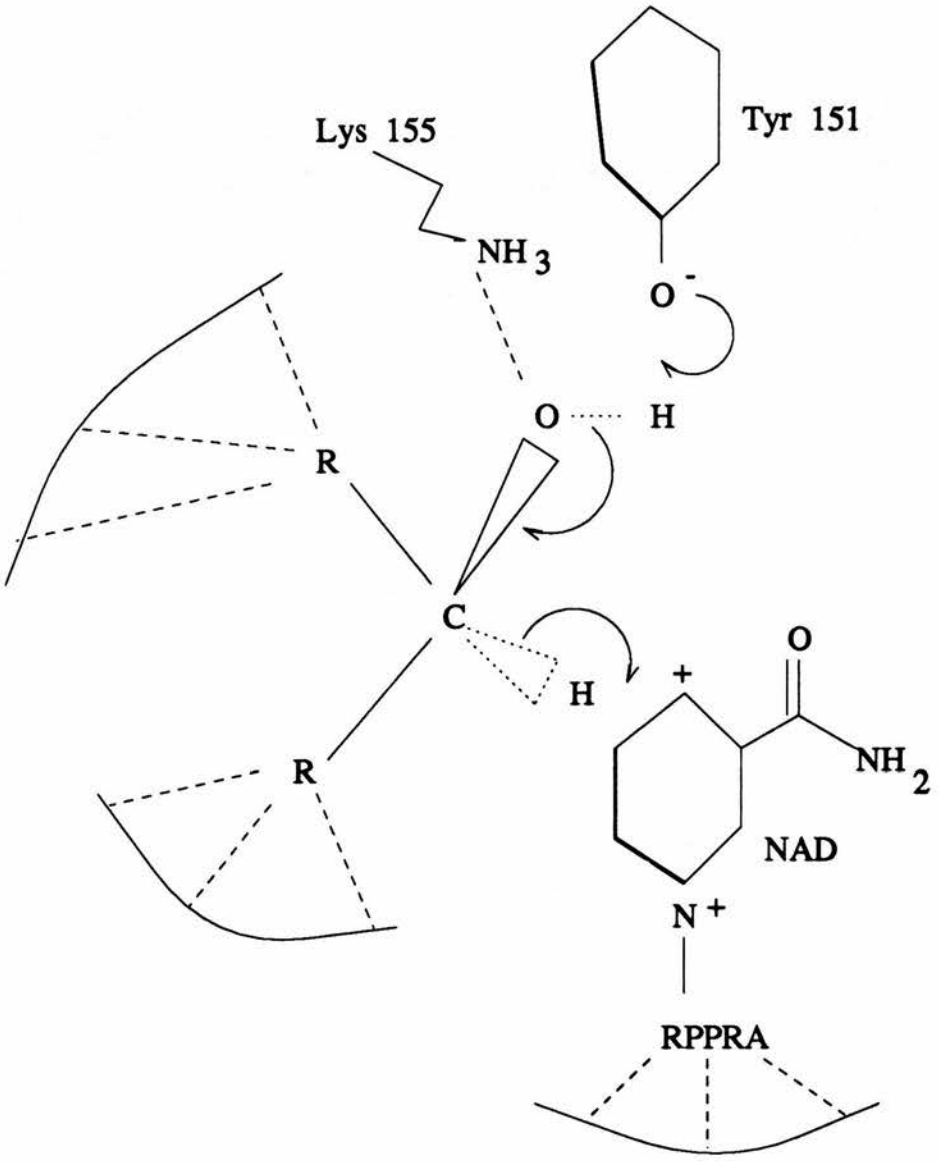


Figure 1–8: Proposed mechanism for DADH based on work by Winberg and McKinley-McKee (1992).

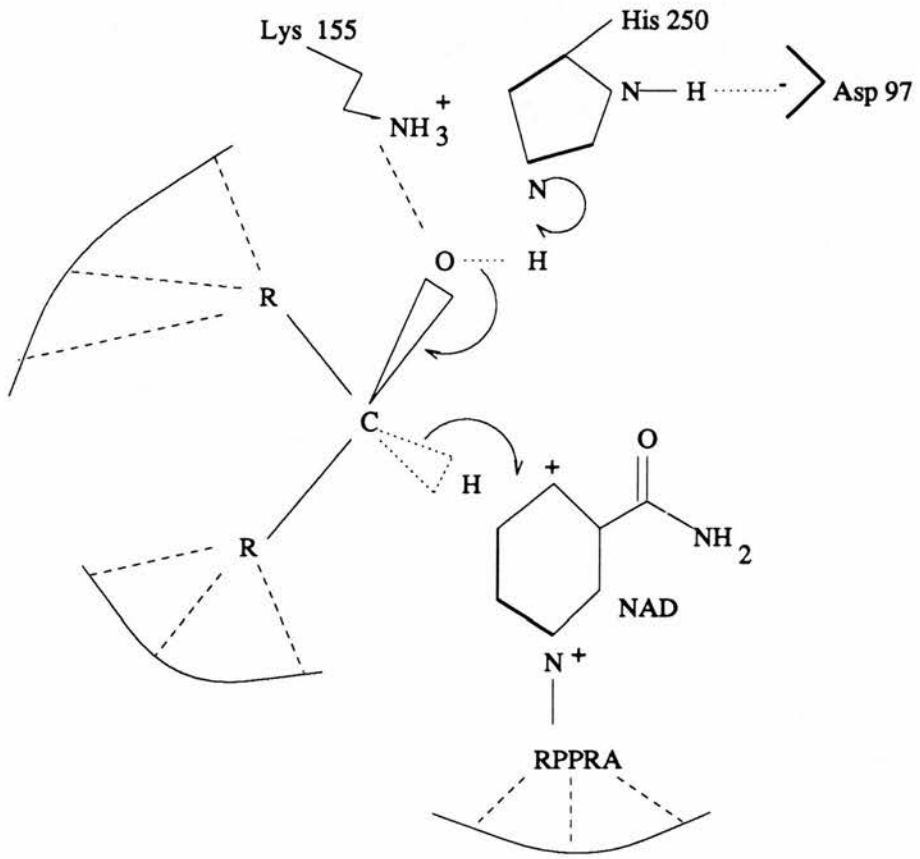


Figure 1–9: Proposed mechanism for DADH as proposed by Ribas de Pouplana and Fothergill-Gilmore (1993).

inactive, is probably due to the effect of this cleavage on NAD(H) binding (Krook *et al.*, 1992).

This mechanism does not account for the RER mechanism proposed by Heinstra (1988) but advocates an ordered reaction mechanism instead. Nor does the unprotonated form of the enzyme group have a negative charge, as suggested by Winberg and McKinley-McKee (1992) which would aid alcoholate ion formation in the ternary complex and the mechanism fails to suggest a role for the conserved Tyr residue, other than as part of the architecture of the active site, even though site-directed mutagenesis of this residue inactivates DmADH (Chen *et al.*, 1992) and other short chain dehydrogenases.

In conclusion, there are probably aspects of each mechanism that are true, but certainly there is not enough evidence to suggest one above the other. Neither, can account for all of the experimental data available for the DADH enzymes.

The evaluation of the proposed mechanisms and topologies for active sites (Chambers, 1991; McKinley-McKee *et al.*, 1991; Winberg and McKinley-McKee, 1992) awaits a structure determination.

1.4.6 Structure

Many sequences for DADH are available. Secondary structure prediction studies indicate that this enzyme has a dinucleotide binding fold which lies in the N-terminal end of the the polypeptide chain. (Thatcher and Sawyer, 1980; Benyajati *et al.*, 1981; Thatcher and Retzio, 1980). A more recent structural analysis of the c.d. spectra predicted 29(2.1)% α helices and 31(2.6)% β strand for ADH from *D. melanogaster* and 28(1.0)% α helices and 41(1.8)% β sheet for the ADH from *D. lebanonensis.* These results confirm the results of Thatcher and Sawyer (1980).

The cDNA analysis showed that the protein encoding section of the DNA was interrupted by two introns, one was located between the codons for amino acid residues 32 and 33 and the second was located between codons for residues 167

and 168. According to secondary structure prediction the first intron lies at the end of the domain predicted to bind the adenine portion of the cofactor. This type of gene arrangement has been observed in the medium chain dehydrogenases (Eklund and Brändèn, 1987) and is consistent with the observation that the dinucleotide fold has been formed by gene-duplication and observations that structurally stable/important domains are coded by single exons (Blake *et al.*, 1983).

Natural mutants of the DADH have indicated that Gly 14, Gln 82, Lys 192 and Pro-214 are all involved, directly or indirectly, with cofactor binding.

Highly conserved residues in the short chain dehydrogenase family indicate that Tyr 152 is conserved in all members of this family, a recently proposed mechanism suggests this Tyr is a possible catalytic residue in the DADH enzyme (if not all short chain dehydrogenases). The active site of the DADH must lie close to the NAD^+ binding site (Winberg and McKinley-McKee, 1988b). Several papers have suggested topologies for the proposed active site of DADH (Winberg *et al.*, 1982; McKinley-McKee, J.S. and Winberg, 1992; Ribas de Pouplana and Fothergill-Gilmore, 1993).

1.5 Scope of this thesis

Crystal structures of enzymes have not only given understanding of protein structure but they have been the most important factor in investigating enzyme mechanism. The crystallographic studies carried out on the horse liver alcohol dehydrogenase (Eklund and Brändén, 1987 and references therein) have revealed much about the enzyme mechanism and how the medium chain dehydrogenases are structurally related.

The aim of the work described in this thesis is to determine the crystal structure of the alcohol dehydrogenase from *Drosophila.* Once a high resolution structure has been obtained then it can be correlated to the vast amount of biochemical information already available for this enzyme. Data from the natural variants

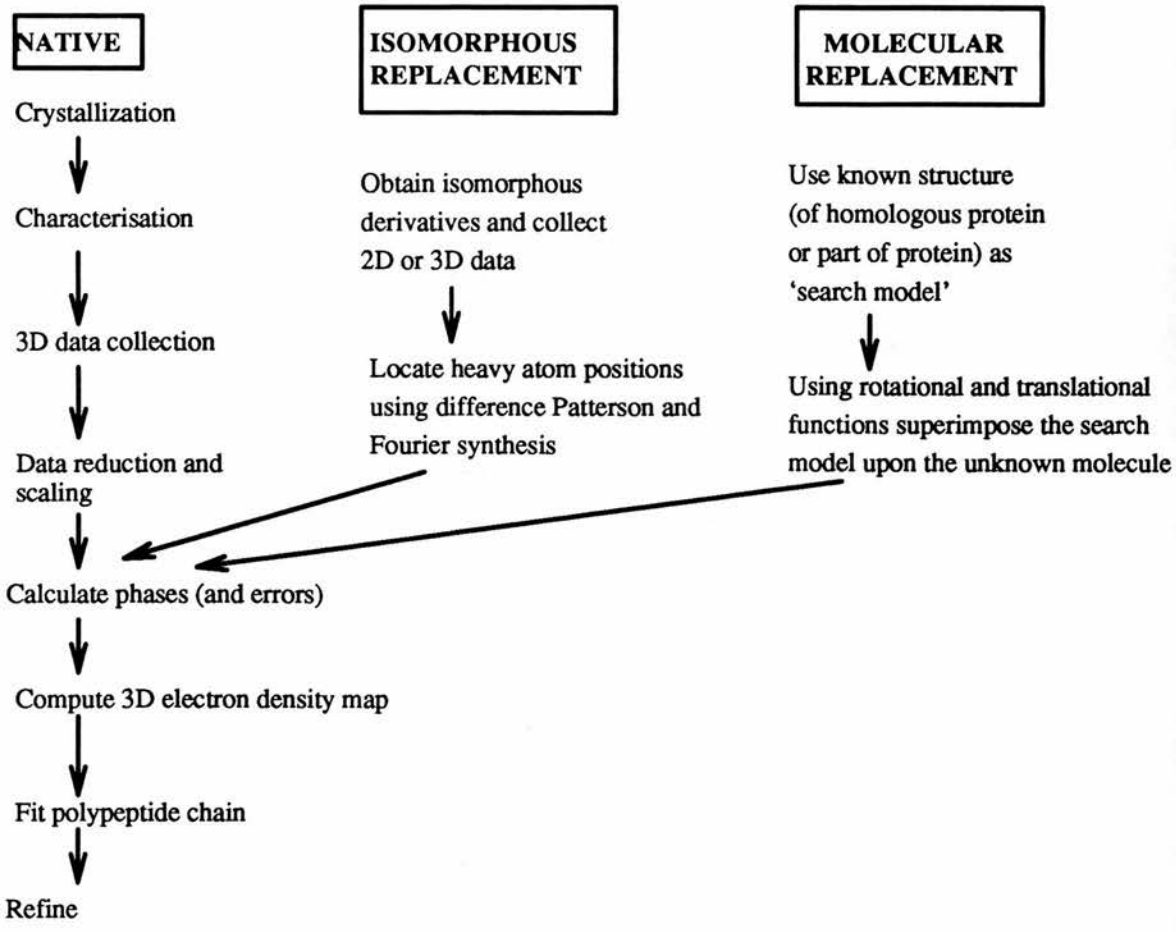


Figure 1–10: Diagram showing the steps involved in protein crystallography, the stages covered in this thesis are detailed in the text

can then be directly correlated with the crystal structure to see why a single mutation can change the biochemical characteristics of the enzyme.

An overview of the steps involved in protein crystallography is shown in figure 1–10.

Chapter 2, gives details of the crystallization of the alcohol dehydrogenase enzyme from *Drosophila lebanonensis*. Two crystal forms were observed. A protocol was developed which produced high quality crystals (form B crystals) of the enzyme. These crystals are suitable for high resolution X-ray diffraction studies. Crystals grown in the presence of imidazole and pyrazole were also grown but data have not been collected on these crystal forms.

Attempts at crystallizing the alcohol dehydrogenase from *Drosophila melanogaster* met with little success.

Chapter 3 describes and summarizes the data that have been collected on both crystal forms of the DlADH. Data collection and the preparation of heavy atom isomorphous replacement derivatives focused on using the form B crystals.

The phasing of the crystallographic data was attempted by both molecular replacement and isomorphous replacement techniques. Chapter 4 gives an account of the molecular replacement study. A polyalanine model of the HSD structure was used as a search model. Several different programs were used and the different solutions were compared and contrasted. Two promising solutions were obtained, but with slightly different orientations. The solution to the translation function was essentially the same in both cases. These solutions were checked using the information obtained from isomorphous replacement experiments.

Chapter 5 covers the preparation and analysis of heavy atom derivatives. The DADH crystals seem to be sensitive to the binding of heavy atom complexes. However, the data from three weakly substituted isomorphous derivatives were processed and the data are discussed.

The initial maps produced using the molecular replacement solution and the isomorphous replacement data were of poor quality and an unambiguous chain trace was not possible. Density modification techniques were employed in an effort to improve the quality of the maps. The resulting maps are described and discussed.

Chapter 7 concludes this study and indicates the direction that future studies might take.

Chapter 2

Crystallization

2.1 Introduction to crystallization

The first step in any crystallographic study is to obtain well ordered crystals that diffract X-rays. To grow protein crystals, it is necessary that there are several milligrams of protein available (say 1-10 mg) and that the protein is as pure as possible and certainly better than 95% pure. This is the minimum quantity necessary when carrying out initial crystallization trials; for a complete structure determination it is likely that much more protein will be required over a period of time.

There are several good reviews covering the crystallization of proteins, that contain more complete discussions; McPherson, (1982) ; McPherson, (1990); Blundell and Johnson (1976); Ducruix and Giegé, (1991); and Weber, (1991). And a collection of recent papers, Carter (1990).

When a solution of protein is brought to a point of supersaturation (this occurs when, for example, solvent is removed by vapour diffusion), as the solution tries to regain equilibrium, solute molecules return to the solid state (see Figure 2-1). However, if molecules in the solid state are absent at this point, the solution will remain supersaturated. There is an energy barrier that requires that the solution moves further into the supersaturated region; as the solution enters this labile region, solid state particles or nuclei are spontaneously formed. If these nuclei are stable (that is the molecular aggregation is coherent so that new molecules are added to the surface more quickly than they are lost back into solution), the nuclei will continue to grow, for as long as the solution remains supersaturated. The rate of crystal growth depends on the distance of the labile state from equilibrium. If the system is pushed too far into this labile region, nucleation occurs in a spontaneous and uncontrollable manner, causing showers of small, mis-shaped crystals. If crystal growth is too rapid then the crystals are prone to flaws and dislocations. The growth of the crystal stops when equilibrium is reached or when the flaws and dislocations in the crystal prevent coherent growth. Since any system aims to maximise entropy and minimise Gibbs free

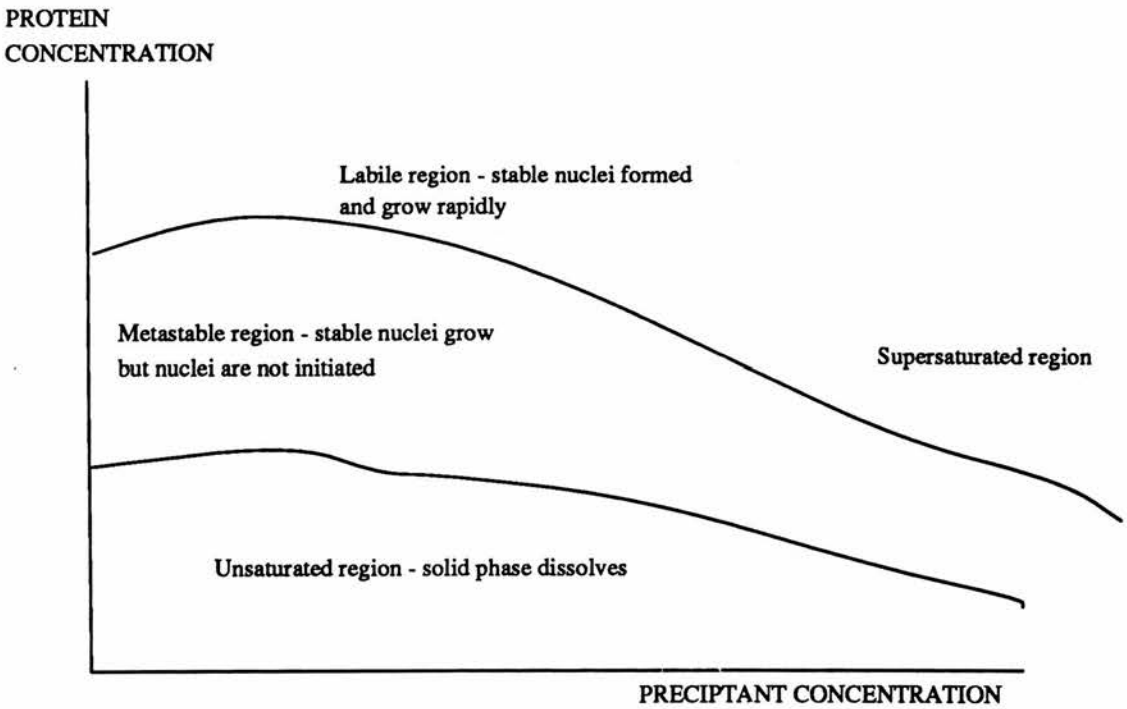


Figure 2–1: A phase diagram showing protein solubility as a function of precipitant concentration.

energy, crystal formation is dominated by the formation of noncovalent chemical and physical bonds, which provide significant negative free energy. A successful crystallization medium contains components that will ensure the greatest number of stable interactions between molecules.

Factors that affect the way in which crystals grow are those that either affect the stability of the protein in solution or that affect the rate of crystalline growth (see Table *2-1*). Proteins themselves are complex and generally labile. By changing their environment the physico-chemical properties of the proteins are changed. For example, the conformation, charge state and size of a protein are all influenced by the surrounding medium. There is a need when crystallizing proteins to use gentle and restrictive techniques that keep the proteins thoroughly hydrated, near physiological pH and temperature. The result of all these restrictions is a multicomponent crystallization system with many variables and as such it can be a complex process to find and optimise conditions that give protein crystals suitable for X-ray diffraction studies.

There are several papers detailing the use of screening methods for crystallization conditions but generally the approach is one of trial and error starting from conditions which have worked for other proteins. (Jancarik and Kim, 1991; Carter, Jr. and Carter, 1979). There are several points to be noted when first attempting to crystallize a protein:

- The factors influencing the stability and activity of the protein should be noted - a crystallographer's aim is to get an atomic structure of an **active** protein.

- Thawing and freezing or lyophilized proteins should be avoided since these processes can destroy protein structure. (However repeated thawing and freezing of a protein has sometimes helped in crystallizing a difficult protein! (McPherson, 1982)).

- Protein concentrations of about 5 $mg.ml^{-1}$ are needed for initial trials, the

Precipitants
Salts: ammonium sulphate, ammonium formate, sodium citrate
PEG: 400, 3000, 4000, 6000, 8000, 20000
Organic solvents: MPD, ethanol (at 4° C; difficult to avoid evaporation)
Mixtures: PEG + 0.5-1M LiCl or NaCl, salts + 2-4% organic solvent
Additives
0.25 - 1% non-ionic detergent, e.g. β-octyl glucoside
Dioxane
Metal ions, e.g. Ca^{2+},Zn^{2+}
Reducing agents: DTT, mercaptoethanol
Glycerol
Azide
Variables
pH
Buffer type
Temperature; commonly tried are 4°, 10°, 17°C
(Gravity, crystallization trials in space)
Co-crystallization partners
Inhibitors
Co-factors
Substrates
Antibodies
Protein variants
Protein purity
Protease digestion
N- or C- terminal truncation
Different species/mutants
Different expression system (glycosylation etc)

Table 2–1: Examples of conditions to change in crystallization trials. Based on Jones and Stuart, 1992

protein should be dissolved in a low ionic strength buffer that does not affect the other components in the crystallization system.

- The buffers used should be stable and soluble at the chosen incubation temperature.

- The number of nucleation points can be limited by using dust free glassware and spinning or filtering solutions to get rid of dust particles and amorphous material from protein solution and buffers.

- Microbial growth in buffers can be prevented by adding trace amounts of sodium azide or thymol.

- Protein and protein crystals are prevented from sticking to crystallization chambers by using siliconised glass and plastic.

There are several techniques employed for growing protein crystals, the most popular is the 'hanging drop' method. A drop (typically 6-14 μl), is formed by mixing protein solution with the solution containing precipitant (50:50). This drop is then suspended above the rest of the precipitant solution which is used as a reservoir. Water leaves the drop until the vapour pressure of the drop is in equilibrium with that of the reservoir solution, this slowly increases the concentration of precipitant in the drop, so that a state of supersaturation is reached. The 'hanging drop' method is suitable for screening a large number of crystallization conditions, while using only small amounts of protein. It can be extended to the sitting drop technique which uses larger volumes of protein solution, which may yield larger crystals. Another popular crystallization technique is that of batch crystallization, in which protein and precipitants are mixed directly. Although this often yields large crystals, it requires accurate knowledge of the crystallization conditions and it uses relatively large quantities of protein and is therefore not suitable when screening for crystallization conditions.

2.1.1 Suitable crystals

A crystal that is suitable for X-ray diffraction analysis should be:

- a single crystal
- 0.2-1.0 mm long in all dimensions for in-house data collection (it is possible to use smaller crystals when using synchrotron radiation)
- able to diffract X-rays, the limit of diffraction depends on the degree of internal order of the crystal

2.2 Crystallization of ADH from *Drosophila*

This chapter describes how stable, highly ordered single crystals of the alcohol dehydrogenase enzyme from *Drosophila lebanonensis* where obtained. Some crystallization trials were also carried out on the alcohol dehydrogenase from *Drosophila melanogaster*. All chemicals were obtained from Sigma unless otherwise stated.

2.2.1 Purification of *D.lebanonensis*

The ADH from *Drosophila lebanonensis* was purified from frozen adult flies, using a modification of the method of Juan and Gonzalez-Duarte (1980) as described by Ribas de Pouplana *et al.*, (1990). Protein used in crystallization trials was passed through an additional sephacryl S-200 column, to ensure protein purity.

The protein (0.3 $mg.ml^{-1}$) was transported from Barcelona, Spain, in 20 mM Tris-HCl, at pH 8.6, with 1% isopropanol, 0.2% mercaptoethanol and 10^{-4} M DL-dithiothreitol (DTT). The buffer had been purged with nitrogen gas. The protein was first dialysed against 20 mM Tris-HCl at pH 8.6 for no more than 24 hours. If necessary the dialysed protein solution was centrifuged using a 50 Ti rotor at 15,000 r.p.m. for 30 minutes at 4°C, to remove any precipitate. About

15 ml of protein solution was then concentrated to 2 ml of solution using an Amicon concentrator, with YM10 filter, under pressure (5 bar) at 4°C. The final concentration was approximately 5 mg.ml^{-1}. The protein concentration was monitored by looking at the absorption at 280 nm (Molar absorbance coefficient, ϵ=13.3 x 10^4 M^{-1}cm^{-1}). At this stage any amorphous precipitant of protein was removed by centrifuging at 7000 r.p.m. for 15 minutes.

2.2.2 Crystallization of ADH from *D.lebanonensis*

The crystallization trials were carried out in 50 mM citrate/Na_2HPO_4 buffer. Initially the hanging drop technique was used to crystallize the protein at 10°C. A range of conditions gave crystals: pH 6.8 - 7.2, with PEG 4000 concentration 14-17% and trace amounts (0.2%) of sodium azide added to all buffers. Crystals were grown in the presence and absence of cofactor (NAD^+).

2.2.3 Crystals

Crystals grew as flat plates in the presence and absence of cofactor (see Figure 2–2). They grew to greater than 0.50 mm in two dimensions, but only 0.05 mm in their third dimension. Crystals grown this way had a shelf-life of only a few weeks, after which time the surfaces of the crystals became pitted, the crystals appeared to begin to redissolve and the diffraction patterns showed disorder. Addition of 1-3% of PEG 4000 failed to halt the degeneration of these crystals.

When diffraction data were collected on these crystals, they were also found to be badly twinned. This twinning had not been observed under a polarising light microscope. The crystals diffracted X-rays weakly and initial characterization was carried out at PX station 9.6 at the synchrotron radiation source (SRS) at Daresbury, U.K. Details of the data collection can be found in chapter 3.

Figure 2–2: Photograph of plate-like crystals or form A crystals of ADH from *Drosophila*.

2.2.4 Refinement of crystallization conditions

Owing to the instability of the crystals and the irregularity of protein supply, efforts were directed to producing more stable crystals. The latent instability of the protein in solution was thought to be the cause of the unstable crystal form. Further crystallization trials were carried out with reducing agent added to all buffers to prevent oxidation of the free sulphydryl groups. Alcohol dehydrogenase from *D.lebanonensis* has two cysteine residues per monomer. Trials with inhibitor present in the crystallization buffers were also carried out.

The buffers were as described above, with the exception that 10^{-4}M DTT, was added to all buffers. Crystallization trials used the sitting drop method, where the total drop size was 20 μl above reservoirs of 0.5 ml. An increase in the amount of protein per drop was used in an attempt to increase the size of the crystals. Drops were placed on siliconised glass pots that were placed upside-down in the culture plates (plastic crystallization bridges (supplied from Crystal microsystems, K. Harlos, 47 Purcell Road, Oxford, OX3 OHB, U.K.) were also used but the crystals stuck firmly to their surface.) Crystallization trials were carried out in the presence and absence of cofactor. A pH range of 6.8 to 7.2 with precipitant concentrations of 14 to 20%, PEG 4000, was tried and a new 'chunky' morphology crystal (form B) was observed growing on those plates without cofactor, within 2 weeks. The largest crystal grew up to 1.0 x 0.8 x 0.8 mm. However, the crystals were more typically 0.25 x 0.25 x 0.25 mm (see Figure 2–4). Some crystals with the old plate-like morphology were observed but generally these appeared in drops where no new morphology crystals grew. Crystals grown in the presence of 0.3 mM NAD^+ were all plate-like.

Seeding

It was found that the size of some of the smaller crystals, whose typical dimensions were less than 0.1 mm, could be improved by using seeding techniques. The small crystals were washed in fresh crystallization buffer where the concentration of precipitating agent was several percent lower than in their

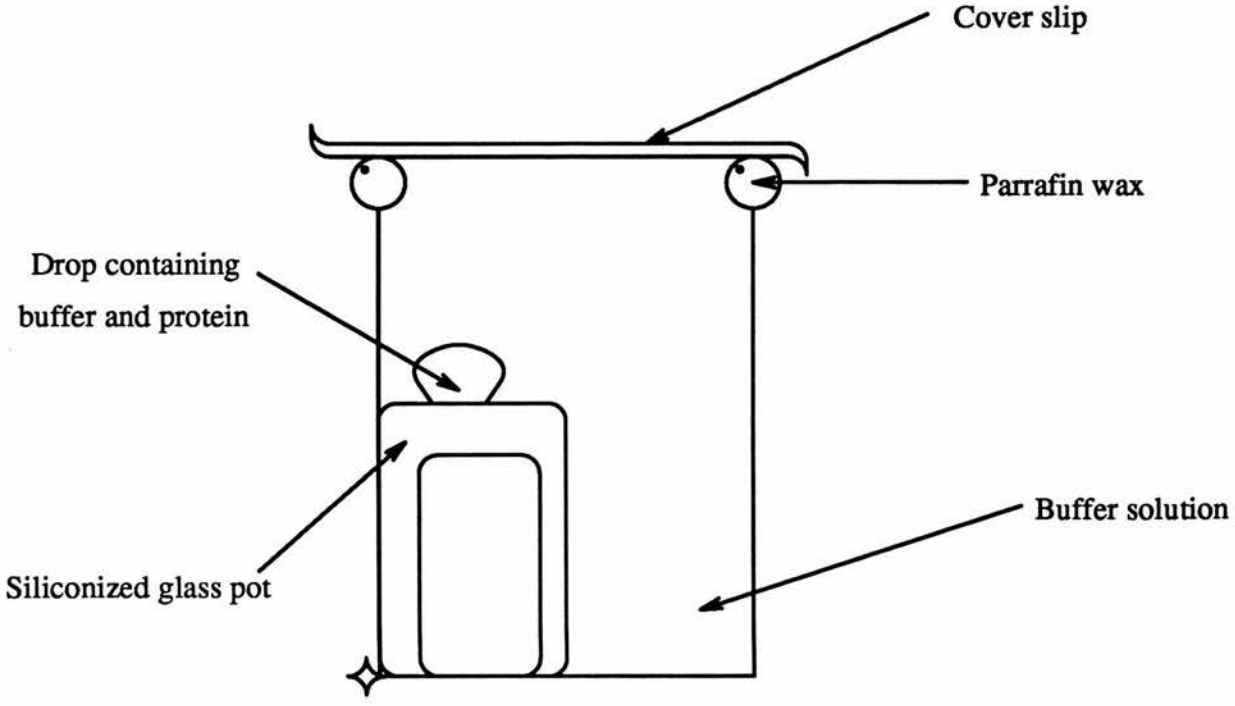


Figure 2–3: Diagram showing the sitting drops used for crystallization

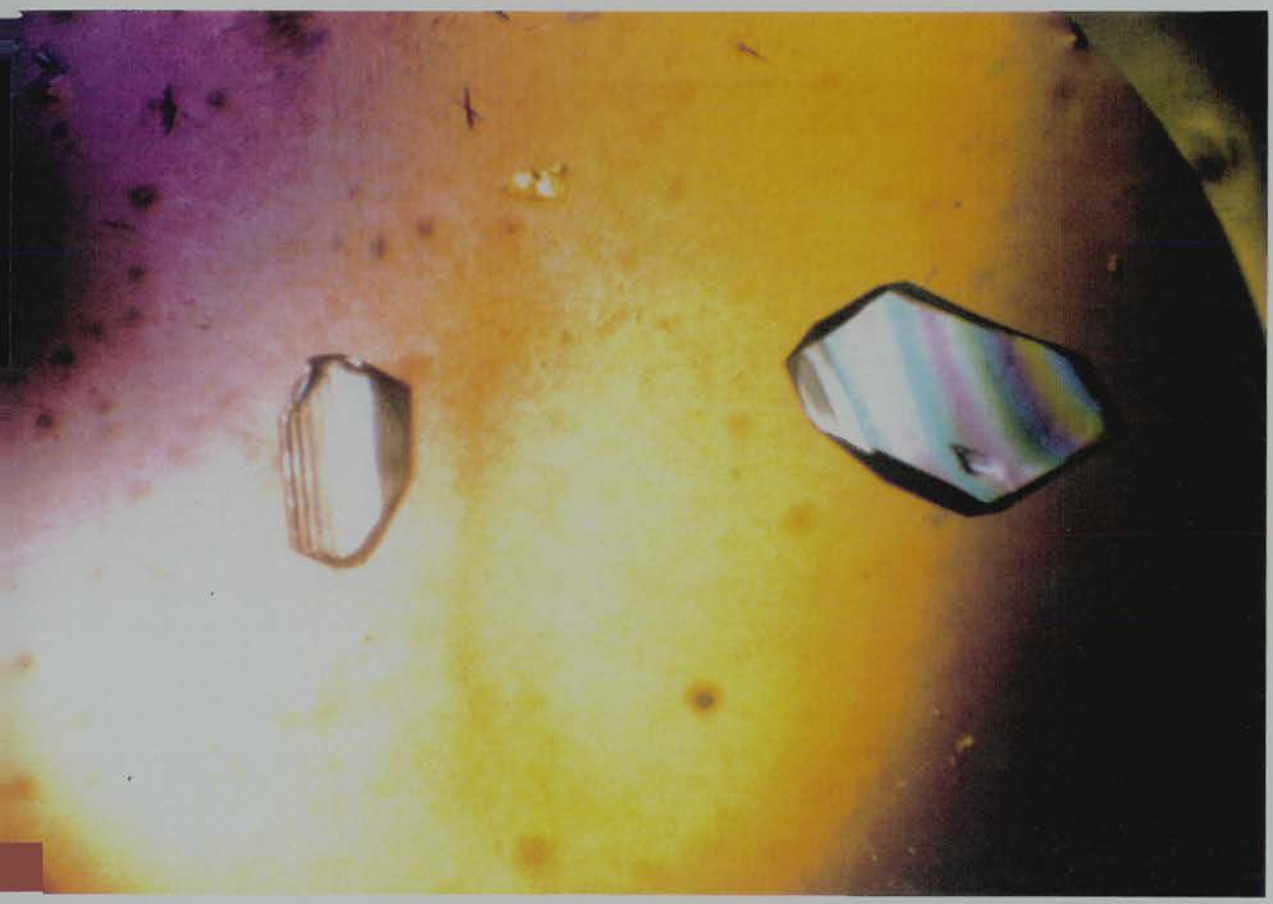


Figure 2–4: Plates showing the new morphology crystals (form B) obtained by adding reducing agent to all crystallization buffers.

mother liquor. This wash removes satellites and other surface impurities that might otherwise act as nuclei in the seeding experiment. A washed, single crystal was then launched into a fresh crystallization drop. The conditions of this drop must be such that they promote crystal growth but not promote nucleation (i.e. precipitant concentration, 14-16% PEG 4000).

2.2.5 Cocrystallization

Trials with crystallization in the presence of inhibitors were attempted. Crystallographic studies of proteins have shown that substrate or cofactor binding often induces some conformational change in the protein where the protein-cofactor complex is often more stable than the native protein. DLADH is unstable in solution. The protein precipitates readily, often before crystallization trials could be set up. In an attempt to stabilize DLADH, cofactors and inhibitors were added to the protein solution and crystallization trials were carried out in the presence of these additives.

Cocrystallization trials of the protein in the presence of cofactor and inhibitor were carried out using the hanging drop method with 6-10μl drops. Crystallization was in 50 mM citrate phosphate buffer (pH 6.9 to 7.1) with a range PEG 4000 concentrations (6, 10, 14, 16, 18 and 20 %); 10^{-4}M DTT and 0.2% sodium azide were added to all buffers.

Crystallization trials of the DlADH in the presence of 3 mM NAD^+ failed to produce form B crystals. Only the unstable form A crystals were formed.

Trials with 2 mM - 3 mM imidazole added to all buffers at higher precipitant concentrations, produced showers of small crystals overnight. The majority of these crystals had the form B morphology. Although, they were less than 0.1 mm in their largest dimension. Further trials, using lower precipitant concentrations and sitting drops in an effort to increase crystal size failed to produce form B crystals, which leads us to suspect that the rate of crystal growth is a factor in determining crystal morphology.

2 mM pyrazole with 0.3 mM NAD^+ present in all crystallization buffers was used with protein reclaimed from an earlier trial. Hanging drop plates were set as described above for the imidazole experiment. A single, new morphology crystal grew after 1 month, at pH 6.8 and 16% PEG 4000 with 3 mM NAD^+. This crystal was approximately 0.1 mm in its longest dimension. An attempt to reproduce this result failed.

Cocrystallization of heavy atom complex inhibitor analogues was attempted by adding 10^{-3}M $KAu(CN)_2$ or $K_2Pt(CN)_4$ to all crystallization buffers (Blundell and Johnson, 1976; also see chapter 5). Badly twinned plate crystals grew with both heavy atom complexes. No other crystal morphologies were observed.

The reproducibility of these trials has not been extensively tested due to the lack of protein. The reproducibility of crystallization trials varies between protein batches. At least in one instance, this was due to a slight modification in the purification step; for example, the protein was passed through the blue-Sepharose column (affinity column) twice. This shifted the crystallization conditions slightly so that crystallization occurred at higher precipitant concentrations.

2.2.6 Purification of *D.melanogaster*

The alcohol dehydrogenase from *D.melanogaster* (DmADH) was purified in the same way as the DlADH.

2.2.7 Crystallization of *D.melanogaster*

Pure ADH from *D.melanogaster* was transported in 20 mM Tris-HCl at pH 8.6, with 1% isopropanol and 10^{-4}M DTT in buffers that had been purged with nitrogen gas. As with the protein from *D.lebanonensis* the protein was dialysed against 20 mM Tris-HCl and then concentrated using an Amicon concentrator, with a YM10 filter.

The crystallization of this protein was approached in two ways:

- Using the crystallization conditions for the *D. lebanonensis*

- Using no prior knowledge of crystallization conditions, a wide range of crystallization conditions and precipitants was tried.

Crystallization using the hanging drop method was carried out in 50 mM citrate phosphate buffer without NAD^+, pH 6.4 -7.4, and with precipitant PEG 4000 8-22% . An amorphous precipitant was obtained at 22% PEG 4000, the protein in the other wells remained in solution. Conditions were changed so that the precipitant concentration in the wells was increased to 15-17% PEG 4000 but crystals failed to grow.

The second approach used several precipitants; ethanol, 2-methyl-2,4-pentanediol (MPD), ammonium sulphate and sodium chloride. However, none of these gave anything other than an amorphous precipitate at the highest concentrations. Other trials were equally unsuccessful, over a range of pH of 3-8 pH units.

The ADH from *D.melanogaster* appears to be more unstable in solution than ADH from *D.lebanonensis*, which may be the reason why it does not crystallize readily. Future studies would recommend that these crystallization trials be carried out with reducing agent in all buffers.

2.3 Discussion

A suitable crystal form of DLADH has been obtained and these crystals diffract X-rays to better than 2 Å resolution. Crystals grow in 50 mM citrate phosphate buffer, pH 6.8-7.2 with 16-22% PEG 4000, with some variation between protein batches. Showers of small crystals observed at higher precipitant concentration, can be used to seed drops at lower precipitant concentrations and hence form usable crystals. Cocrystallization experiments in the presence of $K_2Pt(CN)_4$, $KAu(CN)_2$ and NAD^+ all failed to produce form B crystals. Indeed the presence of these species seem to promote form A crystal formation. It seems likely that the binding of these inhibitors/cofactor induces a conformational change in the

protein which prevents it forming form B crystals. All these species bind to the cofactor binding domain of the protein, it is speculated that in the form B crystals this site forms an important crystal contact. It is also speculated that the oxidation of one or both of the sulphydryls in the monomer has a similar effect on conformation or affects the cofactor binding site in some way since oxidation of the sulphydryls prevents formation of form B crystals. The cysteines in the *D.lebanonensis* monomer lie at positions 137 and 217. From sequence alignments with 3α,20β-hydroxysteroid dehydrogenase and from secondary structure predictions of the *Drosophila* ADH, residue Cys 137 lies at the end of the dinucleotide binding region and Cys 217 lies in a loop region closer to the C-terminus. There is evidence that residue 214 is involved in cofactor binding (Heinstra *et al.*, 1988) which supports the theory that the Cys 217 has an effect on cofactor binding. It is possible that the loop in which Cys 217 lies, folds back to lie close to the cofactor binding domain. If this proves to be the case then it can be seen how the oxidation of one or both of these cysteine residues could induce a change in the protein which affects conformation/accessibility of the dinucleotide binding domain and hence crystal morphology. We are assuming that the two morphologies represent different crystal forms and not just different crystal habits (see Chapter 3). A recent study has highlighted that oxidation of the sulphydryl groups decreases the stability of the DmADH in solution.

Why does *D.melanogaster* fail to crystallize ? The limited attempt made to grow crystals of *D. melanogaster* makes it difficult to conclude a great deal about the crystallization of this protein. However, we have since become aware that extensive crystallization trials of *D. melanogaster* had been carried out by G. Chambers (personal communication) several years previous to this study. He had eventually succeeded in growing crystals of the enzyme-NAD-isopropanol tertiary complex. He believes that his lack of success was due to the instability of the *D.melanogaster* enzyme in solution. It appears that success was only achieved when a homogeneous solution of the enzyme-NAD-isopropanol was available. This casts some doubt on the attempts at cocrystallization made in these studies, the addition of NAD^+ to all buffers may produce several different species

of ADH dimer (e.g. with a single molecule of NAD^+ bound, with two molecules of NAD^+ bound or with no NAD^+). Such a mixture of species might lead to problems in crystallization.

Our degree of success in obtaining crystals suitable for diffraction work has been attributed to using ADH from *D.lebanonensis* instead of from *D.melanogaster*. Although less is known about the biochemistry of ADH from*D.lebanonensis*, it is known that the DlADH protein is more stable in solution (Gonzalez-Duarte, personal communication). There are slight differences in the rate of catalysis between DlADH and DmADH, but that the overall substrate specificity has been conserved between the two enzymes and there is evidence that the two proteins have a high structure similarity.

Chapter 3

Data Collection

3.1 Introduction

3.1.1 Data collection

This chapter summarizes details of the data collected from crystals of ADH from *Drosophila*. Included is a brief description of the data collection procedures and reduction programs used.

Data were collected using electronic area detectors and an image plate. These detectors combine the efficiency of the rotation method (Arndt and Wonacott, 1977) with more convenient image processing. Data have been collected using a Xentronics detector (Durbin *et al.*, 1986) an Enraf-Nonius FAST T.V. detector (Arndt and Gilmore, 1979) and a MarResearch Image plate. The data collection strategies employed in each case are described.

3.1.2 General strategy for data collection

The aim of any crystallographic data collection is to collect as much of the unique reflection data as possible. It is also desirable that several recordings of the same or equivalent reflections are measured, since this improves the accuracy of the data and aids internal scaling of the data. The data collection strategy adopted will depend on the symmetry of the crystal, the sensitivity of the crystals to radiation damage and the type of data required. The crystal is placed in an X-ray beam and then rotated about an axis (see Figure 3–1). As the crystal is rotated the Bragg planes are brought into the reflecting position and diffraction can be observed. An image of the diffraction pattern is recorded every few degrees as the crystal is oscillated so that the Bragg planes are brought into the reflection position several times per image. The total rotation range is chosen in order to collect the unique reflection data for a crystal with adequate redundancy.

3.1.3 X-ray source

Two X-ray sources were employed in the course of this study: data were collected on a rotating anode X-ray generator and at the Synchrotron Radiation source (SRS) at Daresbury, U.K. Synchrotron radiation is an intense, highly collimated beam, at least 1000 times more intense than the beam produced by a rotating anode generator (SERC Daresbury laboratory, 1991). This means that data collection at the SRS is much quicker. Also the wavelength is tunable and so can be tuned to the absorption edge of a substituted heavy atom thereby increasing the anomalous signal. Synchrotron radiation is used when crystals are very small, or diffract X-rays weakly or when the high resolution weak data is required. At short wavelengths, obtainable at the SRS, protein crystals have a low absorbance which means that the radiation damage to the crystals is reduced. Therefore, data can be collected on radiation sensitive crystals.

For the majority of protein crystallographic studies, data to 2.5Å can be collected using a rotating anode source.

3.1.4 Detectors and methods

As mentioned in the introduction to this chapter, several different area detectors were used to collect data, the subtleties of the methods employed in each case are described below.

Image Plate data collection

The rotation method, where the crystal is rotated or oscillated through 1-2° per image, was used for data collection with X-ray film. The technique has recently been revived with the advent of image plate systems (IP). The IP is sufficiently large (90 mm radius or larger) so that it remains in a symmetrical position about the beam stop. In this way positive and negative 2θ data can be collected on the same image. The oscillation range per image is limited to avoid overlapping reflections; which depend on the unit cell dimensions and on the resolution

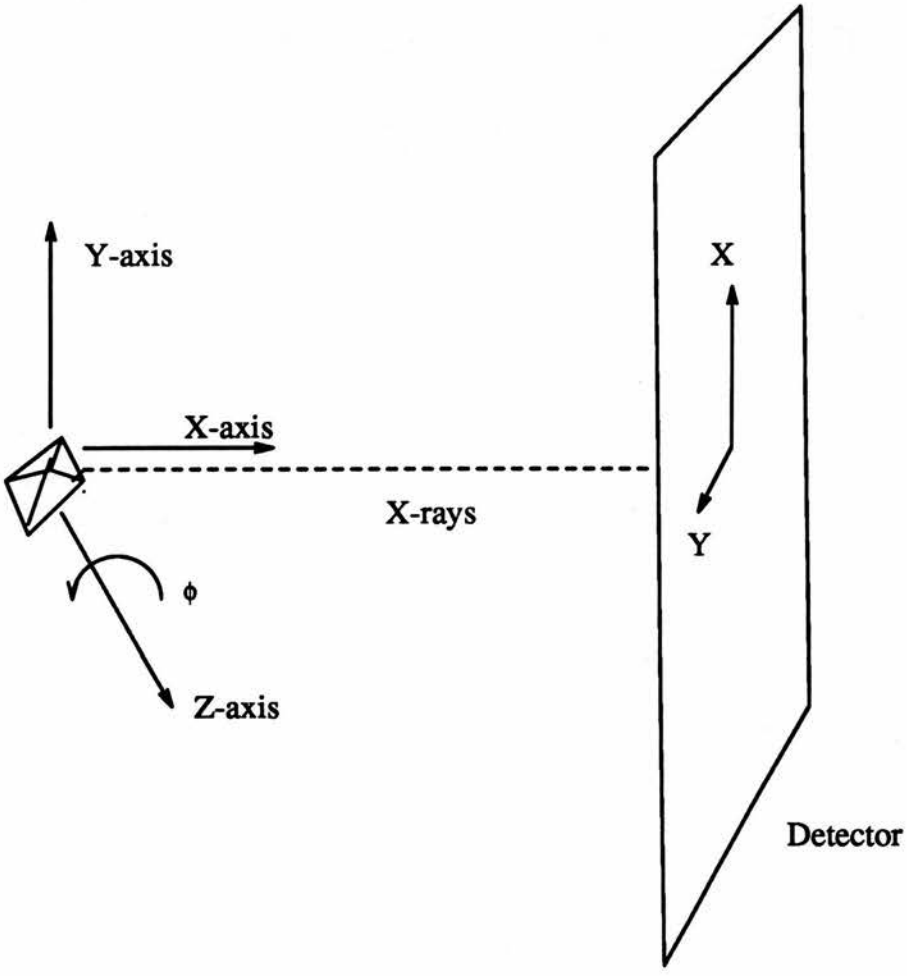


Figure 3–1: Experimental setup for data collection using the rotation method.

range. The total rotation range needed to collect a full data set for the crystal is determined by the symmetry of the crystal. When collecting data using an IP some of the following points should be noted:

- Rotation should be about the longest cell axis so as to maximize the rotation angle per image. This rotation range may need to be changed as the crystal orientation is changed, and other cell axes move to become parallel with the incident X-ray beam.

- Rotation about the axis of highest symmetry minimizes the total rotation range needed to collect the complete data set.

- The signal to noise ratio is poor when 1-2° oscillations are used, but it can be improved by increasing the crystal to IP distance. However increasing the crystal to detector distance limits the resolution of the data. The signal to noise ratio for an image can be increased by reducing the diameter of the collimator, so thus the background intensity of the image is reduced.

- Since the spatial resolution of an IP detector is good, the IP can be moved close to the crystal to collect high resolution data

- For oscillation images there is an error associated with the determination of ϕ (where ϕ is the angle of rotation). However, ϕ must be known precisely before an accurate orientation matrix can be determined. It is therefore good practice before the onset of IP data collection to collect two 'still' images (where the crystal is not oscillated) separated by 90° in ϕ, to be used in refining the orientation matrix.

- If the anomalous signal is to be collected the crystal should be orientated so that the Bijvoet pairs are collected on the same or adjacent images

Electronic area detectors

Electronic detectors use the rotation method and data are produced in a digital form that requires no post exposure scanning. Therefore, collection of successive images is quicker than for an IP. The two electronic detectors used in this study were the FAST and the Xentronics detector.

The FAST is a T.V. based detector which has a phosphor detector deposited on a fibre optics screen. The fibre optics is coupled to an image intensifier which is in turn coupled to a Silicon Intensifier Target (SIT) T.V. camera tube (Schierbeek, 1991). This detector has a high count rate and can therefore be used with a synchrotron radiation source.

The Xentronics detector is a xenon-filled, focused imaging, proportional counter. The incoming X-ray quanta are detected as 'pulses' and stored as they arrive. It has a relatively slow count rate and is only used on a rotating anode source.

Generally, electronic detectors are of limited size and spatial resolution; collection of all data requires that the detector is swung away from a symmetrical position about the backstop, and that the data are collected using several scans and detector positions. The following points should be noted when collecting data on an electronic detector:

- The rotation angle per image is small, only 0.1 - 0.25°, exposures per image are shorter and therefore the images have an improved signal-to-noise.

- Crystal rotation should be around one of the longer cell axes, the shortest axis should be parallel to the incident X-ray beam. When the shortest axis is perpendicular to the incident X-ray beam, reflection spots are close together across the face of the detector and the poor spatial resolution of the detector requires that the crystal to detector distance is increased.

- Rotation about the crystal axis of highest symmetry reduces the total rotation range needed.

- For each reflection, coordinates for X, Y and ϕ are available (where X and Y are the positions on the detector surface and ϕ is the angle of rotation of the crystal (see Figure 3–1)) because there is a better estimate for ϕ. These allow for exact centroids of each reflection to be determined which greatly facilitates autoindexing.

- Partial reflections (reflections that occur on more than one image) are avoided since all reflections are summed over a number of consecutive images.

- If Bijvoet pairs are to be collected, the crystal should be aligned so that as many as possible of these reflections are collected from the same or adjacent images.

IP *vs.* electronic detectors

Image plates are attractive detectors because they have a high sensitivity, a very low background signal, a good spatial resolution and a large detector surface. Therefore, they are ideal for collecting data on crystals with large unit cells, weak diffraction intensities and high resolution data. The quick response time of the IP to incoming photons, makes them useful for data collection at a synchrotron. However, the nature of the autoindexing, the inaccuracy in the determination of ϕ, means that it is not robust, so accurate cell dimensions should be known before data collection; there is much work being carried out developing software for IP data processing. In contrast, the area detectors have good autoindexing routines and data reduction software. However, area detectors have a low spatial resolution, a small dynamic range and a small active detector surface. They are also possibly, less robust and more prone to damage than the IP's.

General strategy for data collection

For spacegroup $P2_1$, the unique data can be collected with a total rotation range of 180°, and collection of cusp data. This rotation range can be reduced to 90° if

the crystal is rotated about the **b** axis provided that the detector is placed symmetrically about the beam stop. In general, 180° of data were collected to give good redundancy for scaling. It is often better to have the crystal slightly misset so that data missed due to the rotation method do not all lie along a single axis.

3.1.5 Overview of data processing

The following list indicates the steps involved in processing macromolecular crystallographic data from an area detector:

- Detector calibration - nonuniformity/floodfield and brass plate
- Autoindexing - determination of crystal orientation and unit cell dimensions
- Prediction of reflection positions
- Integration of reflections
- Postrefinement of orientation matrix
- Data reduction - scaling and merging

Detector calibration

The detectors are first calibrated to correct for the nonuniformity of their response across the detector surface. For the IP detector, this correction is due to errors in the nonuniformity of the scanning mechanism; in electronic detectors it is due to the nonuniformity in pixel response.

The electronic area detectors must be further calibrated to give a distance to pixel conversion. This is done by attaching a brass plate, which has been drilled with a regular array of holes, to the front of the detector. The plate is then

illuminated with an ^{55}Fe source and a distance/pixel calibrated can be done since the distance of separation of the excited pixels is known.

Autoindexing

Autoindexing is used to find the correct orientation and dimensions of the unit cell. It is then trivial to predict the positions of reflections. The orientation matrix is refined throughout the data integration process, this corrects for any small change which might occur. There are several different autoindexing routines:

- IDXREF in XDS written by Kabsch(1988)
- REFIX in IP processing
- IDXREF in IP processing (refinement of the orientation matrix)
- AUTI in MADNES, written by P.Tucker
- AUTJ in MADNES, written by J.Pflugrath

Autoindexing routines, are provided with a set of strong diffraction spots. The most sophisticated and powerful routines calculate difference vectors for these spots and then fit three non-coplanar vectors, to the short difference vectors. These non-coplanar vectors form a basis set of vectors from which the cell parameters and orientation are determined. The cell and orientation parameters are then refined using the longer difference vectors. Most autoindexing routines use strong diffraction spots, that are widely separated in ϕ, to calculate difference vectors. The routines are sturdy and can autoindex with no prior knowledge of cell parameters. Difficulties are encountered if the crystal is twinned or split, or if two unit cell edges are almost the same length.

The autoindexing is slightly different for film and IP's, where an uncertainty in ϕ means that there is an uncertainty in the centroid of the spot position. This introduces a degree of error into the indexing. IDXREF in XDS (Kabsch, 1988)

and REFIX (prior to processing with MOSFLM) identify only two non-coplanar vectors (instead of three non-coplanar vectors, see above). These methods are less robust and presume that the spacegroup and approximate cell dimensions are known. The indexed cell is refined (IDXREF for IP data) using the partial reflections, such that an accurate matrix for the orientation and cell parameters is known before reflection prediction and integration.

Profile fitting

Profile fitting (Diamond, 1969; Ford, 1974 and Rossmann, 1979) is used to improve the accuracy of weak reflections. For weak reflections, the background dominates the total variance, and this leads to inaccuracies when a simple summation integration procedure is used to evaluate the intensities of these weak reflections. Profile fitting assumes that strong and weak reflections share the same intensity profile: an average or standard reflection profile is calculated for a group of reflections and this is then used to obtain a more accurate estimate of the intensity of the weak reflections. Profile fitting is more prone to systematic errors than the summation integration intensities (see later), but generally, profile fitting improves the variance of a reflection intensity.

Rossmann (1979) derived a standard two dimensional profile for processing film data. Two dimensional profile fitting is used when the rotation method is used with large, 1-2° oscillations; for data from an electronic area detector each reflection appears on several consecutive images and therefore a standard three-dimensional profile can be determined. Rossmann fitted a 'shoe box' around a pre-calculated reflection position, the counts per pixel for this 'shoe box' are stored and this forms an intensity profile. A similar method is used for IP data processing (Leslie, 1990). Systematic errors in profile fitting occur in two ways:

- The standard profile, which is averaged over many profiles will be artificially broadened, since the 'shoe box' placed over the reflection is centred on the pixel closest to the predicted reflection position and not

necessarily on the reflection itself. This broadening will lead to on overestimatation of intensity (Diamond, 1969).

- There will be a displacement of the centre of the standard profile with the centre of the reflection profile, this will lead to an underestimation of intensity.

The combination of these effects can lead to an error of measured intensity of 2%, which is significant for strong reflections. This error can be reduced by interpolating the observed counts onto a grid of pixels centred on the predicted reflection position. This improves accuracy if the reflection positions can be calculated to within half a pixel.

Further errors in this method are due to the variation of the reflection profile across the face of the detector. For example, reflections situated away from the rotation axis hit the detector at a more oblique angle than those reflections close to the crystal rotation axis (this is a direct consequence of the rotation method). Rossmann (1979), overcame this problem by calculating a standard reflection profile for a number of separate regions of the detector. An alternative solution to the problem of distortions of the reflection profile over different regions of the detector surface is implemented in XDS and in MADNES where three dimensional profiles are used. In XDS, the standard profile is built using a 'pixel-labeling routine' (Kabsch, 1988b). Each pixel in the image is labeled by the indices of the nearest calculated reflection. Profiles are evaluated on nine areas of the detector surface and repeated each 10° in ϕ . The coordinates of this reflection profile are then transformed so that each reflection profile is represented in an undistorted way, that is, it corrects for distortions arising from the rotation method itself. MADNES implements another profile fitting algorithm written by P.Brick, which also minimizes the distortion of the reflection profiles but it does this by taking into account the path of the reflection through the Ewald sphere, without using the complete transformation into reciprocal space that Kabsch (1988b) uses.

Even with these modified profile fitting routines it has been found that the

internal agreement between strong reflections can be poor, this is thought to be due to systematic errors in the profile fitting for strong reflections (Leslie, 1991). The CCP4 program AGROVATA deals with this problem by calculating a weighted mean of the profile fitted and the summation integration intensity of each reflection. It weights the profile fitted intensity for weak, high resolution reflections and the summation intensity for the strong, low resolution reflections.

Postrefinement

During the course of data processing, the orientation matrix should be continually refined and updated. So that the prediction of reflection positions is accurate throughout the processing. This refinement corrects for crystal slipping (it can also be used to monitor any change in the unit cell dimensions of the crystal that might be due to radiation damage, although generally unit cell dimensions are held constant). Postrefinement for film and IP uses only the partially recorded reflections, since the determination of ϕ for these reflections is more precise than for fully recorded reflections. Postrefinement for electronic area detectors uses strong, well measured reflections for refinement.

Scaling

Once data have been integrated and estimates of their intensities are known, the data must be put on a common scale and reduced to a symmetry-unique set of reflections. Data sets are scaled in an attempt to remove any systematic errors introduced during the data collection and reduction, and also to correct for absorption by the crystal and radiation damage of the crystal during data collection. Some scaling is done within the data reduction programs (XDS, MADNES and ABSCALE (part of MOSFLM)) e.g. multiplication by the Lorentz factor and correction for polarization effects.

Changes in the intensity due to absorption by the crystal are generally not corrected for (as they are in small molecule crystallography) since protein

crystals absorb X-rays weakly and at short wavelengths (λ=0.89Å), absorption effects are reduced even further.

When the reflections are reduced to a standard asymmetric unit, the file may contain several observations of equivalent reflections. The data are then batch scaled; data integrated from different images are put into batches of 5°. The mean intensity of this batch is then scaled to the mean intensity of all other batches (this is a throw back to film data processing where separate films were scaled together by this batch method, this is the way IP images are scaled). In this way all observations are then on the same scale as all other observations. During data collection it is advisable to rotate the crystal about more than one axis since this gives more observations per reflection, which helps when scaling, and fills in the cusp region.

The scale, as determined by ROTAVATA and AGROVATA, is applied as a scale factor, K_j and a relative temperature factor $e^{-2B\sin^2\theta/\lambda^2}$. The temperature factor is mainly to account for radiation damage in the high resolution reflections. This term becomes significant at resolutions greater than 2.5Å . The disadvantage of this type of scaling however, is that the scale factors are independent of each other, and therefore neighbouring reflections can have very different corrections applied if they are not a member of the same batch.

An alternative scaling program, XSCALE (Kabsch, 1988b), determines scale factors on a grid of positions on the detector and as a function of crystal rotation. It also uses an adaptation of the Fox and Holmes (1966) method. XSCALE uses a weighting scheme to determine the scale factor, where the weight applied depends on the distance between each reflection and the scale factor; in this way each reflection contributes to several scaling factors. This scale factor is a continuous function.

3.2 Methods and results

The next section covers the data collection carried out on the DADH crystals, form A and form B. The section provides a roughly chronological account of the data collected on the ADH crystals.

3.2.1 FAST

Initially, characterization of the form A crystals was carried out at the synchrotron at Daresbury, U.K. These crystals diffracted X-rays weakly and appeared to be extremely sensitive to radiation damage. Data were collected at PX station 9.6, wavelength 0.89Å using the FAST area detector. 180° of data were collected as 0.1° images. The autoindexing routine AUTI (which requires no knowledge of the unit cell parameters prior to indexing), was used to index the unknown unit cell. Accurate parameters for the beam and detector positions were input and indexing was carried out, using approximately 70 reflections from two batches of data widely separated in ϕ. The autoindexing succeeded but the solution failed to refine. Attempts to refine this cell (and others) were carried out in AUTJ (which needs some knowledge of the unit cell parameters before indexing and refinement are undertaken). It was only after superimposing 10-20 FAST images that the origin of the problem was identified as being due to crystal twinning . The pseudo-oscillation picture of a form A crystal (see Figure 3–2), clearly shows this twinning which was difficult to identify optically (see Figure 2–2). The crystals appeared to be twinned such that the twins share one axis. This type of twinning is quite common in monoclinic crystals, where twinning is generally found along an axis or a crystal edge perpendicular to the diad axis. Approximate cell dimensions for the twinned crystal were a = 70.0 Å, b = 55.8 Å, c = 150.0 Å and $\beta = 110°$. In the light of later results it appears that the twinning results in a doubling of the **c** axis.

This twinning highlights a problem with data collection using an electronic

Data	Cell				R_{sym} (%)
Native	a = 81.4	b = 56.3	c = 111.3 Å	β=94.8°	6.7
$Pt(NH_3)_2Cl_2$ soak	a = 82.6	b = 56.3	c = 111.6Å	β=95.1°	16.2
$Pt(NH_3)_4Cl_2.H_20$ soak	a = 81.3	b = 55.5	c = 111.2 Å	β=96.2°	~40
$NaIrCl_6$ soak	a = 81.4	b = 56.3	c = 115.4	β=96.0°	

Table 3–1: Table summarising the data collected on the FAST detector at the synchrotron, Daresbury, U.K. R_{sym} is defined as $\frac{\Sigma_{hkl}\Sigma_i^N|\bar{I}-I_i|}{\Sigma N\bar{I}}$.

detector which is, that flaws in the crystal lattice are not immediately apparent as the images produced by the detectors show a limited area of reciprocal space. It is therefore good practice when about to collect data on an electronic detector to look at the diffraction pattern of the crystal at several different rotations to ascertain that there are no double spots, badly shaped spots or any extra symmetry.

Initial characterization of the form B crystals were also carried out at the SRS at Daresbury, PX station 9.6 using the FAST area detector. The crystals were small, typically 0.15 mm in their longest dimension and generally the data were weak and difficult to process. The data were processed using MADNES with EVAL 3 profile fitting (Kabsch, 1988) and also EVAL 6 profile fitting, therein. (MADNES has several profile fitting routines, labeled EVAL 1 - EVAL 6, the label associated with different routines varies depending upon the version of the program used). Autoindexing of the FAST data revealed a cell, a = 81.24(6) Å, b = 55.75(4) Å, c = 109.60(7) Å and β = 94.26(9)°, spacegroup $P2_1$. The large cell has a volume of 495,021 $Å^3$; which gives a V_M of 2.21 $Å^3$/dalton (Matthews, 1968); corresponding to two DlADH dimers in the asymmetric unit. The DlADH crystals have a solvent content of 44% . The data collected on the FAST were eventually discarded due to incompleteness of native data and also because larger form B crystals were grown and it became more convenient to collect data using a Xentronics detector.

Table 3–1 summarizes the data collected on the FAST detector, including some data collected on heavy atom soaked crystals that had been prepared by soaking for 6 hours in 1 mM solutions of heavy atoms compounds listed.

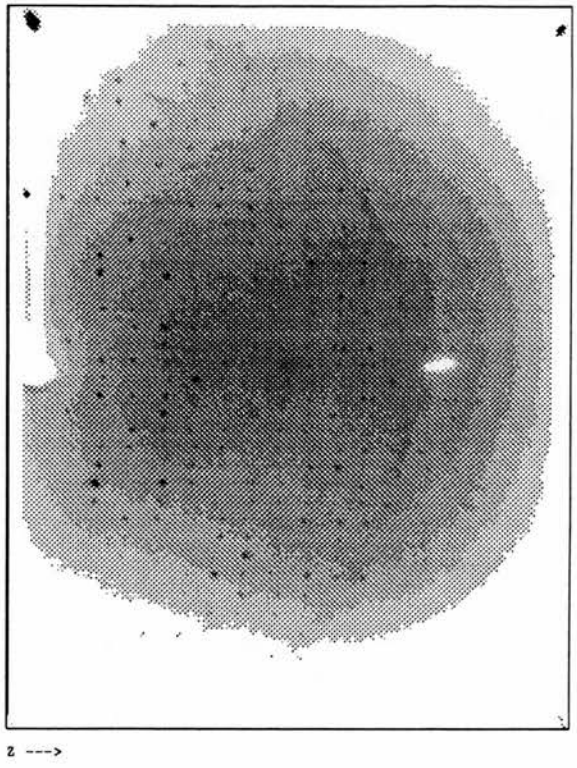


Figure 3–2: Pseudo oscillation picture constructed from 10 images from the FAST detector. The picture shows a twinned lattice.

3.2.2 Xentronics

Once the crystallization conditions had been refined to give a new morphology crystal (form B crystals), data collection became a very much easier task. The new crystal form diffracted well and data could be collected on a lab X-ray source using a Xentronics detector.

However, initial autoindexing of data collected on a Xentronics detector, using IDXREF (a subroutine within XDS) indexed a cell a = 70.6 Å, b = 55.75 Å, c = 65.75 Å and β = 106.95° which is smaller than the cell found using the FAST detector. The larger cell was only found using IDXREF when more data frames, widely separated in ϕ were indexed. The small cell is related to the large cell by a rotation about the **b**-axis (see Figure 3–3). It contains one dimer in the asymmetric unit, compared to the large cell which has two dimers per asymmetric unit. It is thought that the small cell is valid at low resolution but that at high resolution, weak reflections with $h+l$ odd are present, and the large cell is valid. The native Patterson for the large cell shows a strong peak on the Harker section at x = 0.5 and z = 0.5, confirming the pseudo **B**-face centring.

The existence of the small cell was confirmed by taking some small angle precession pictures from one of the DLADH crystals. These pictures (Figures 3–4 and 3–5) confirm the spacegroup as being $P2_1$. However, they have a resolution limit of 4 Å (10° precession, 20 h exposure, CuK_α, 40 mV, 30 mA) so the existence of the large cell at higher resolution is not negated. This exercise also helped to identify the unit cell axes relative to the crystal morphology (see Figure 3–6).

Data were processed using the large cell and then analysis was carried out on the $h+l$ = odd reflections. The intensity of these reflections as a function of resolution is given in table 3–2. The analysis of the $h + l$ = odd reflections shows that on average these reflections are very weak, probably below the noise level. Although some reflections must be significant because autoindexing finds the large cell. The intensity of the $h + l$ odd reflections in the CMN data increases

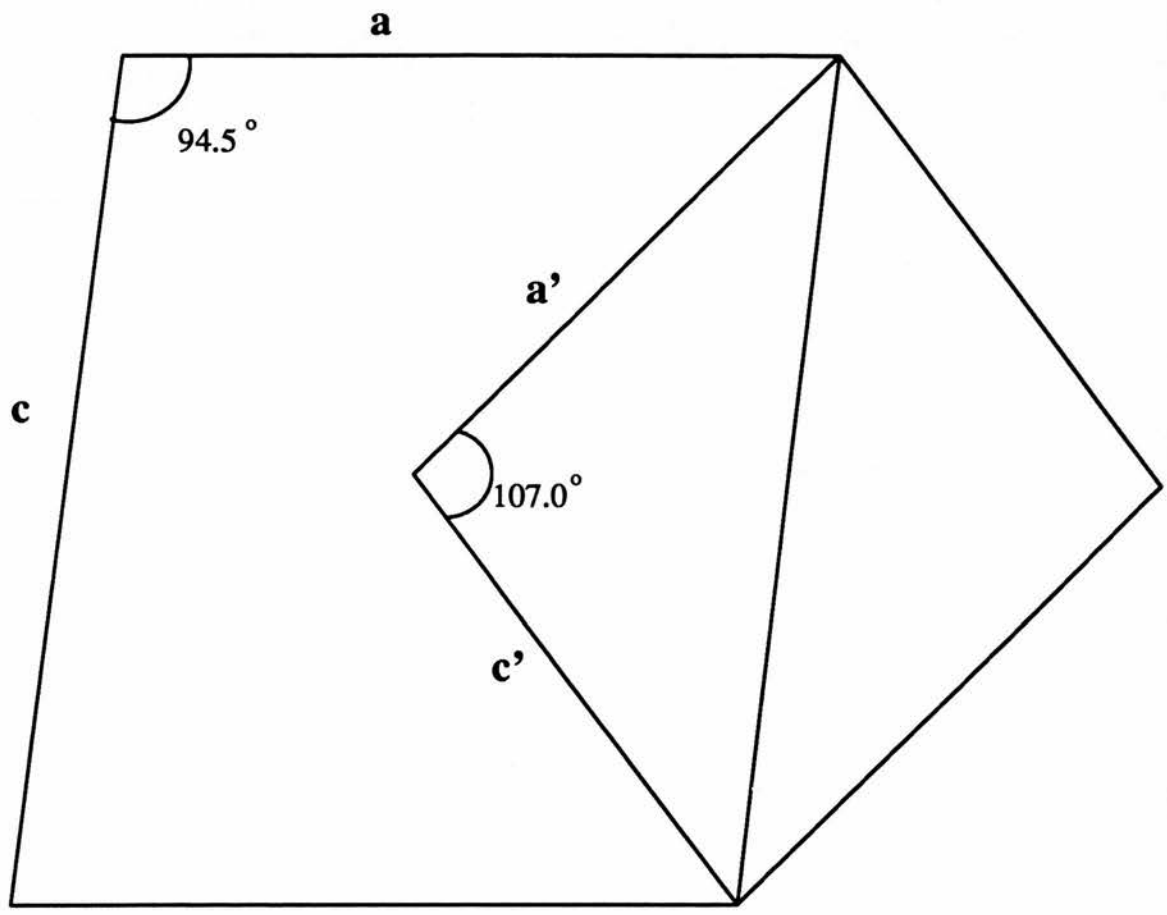


Figure 3–3: Relationship between the large and small unit cell found for form B crystals of DlADH.

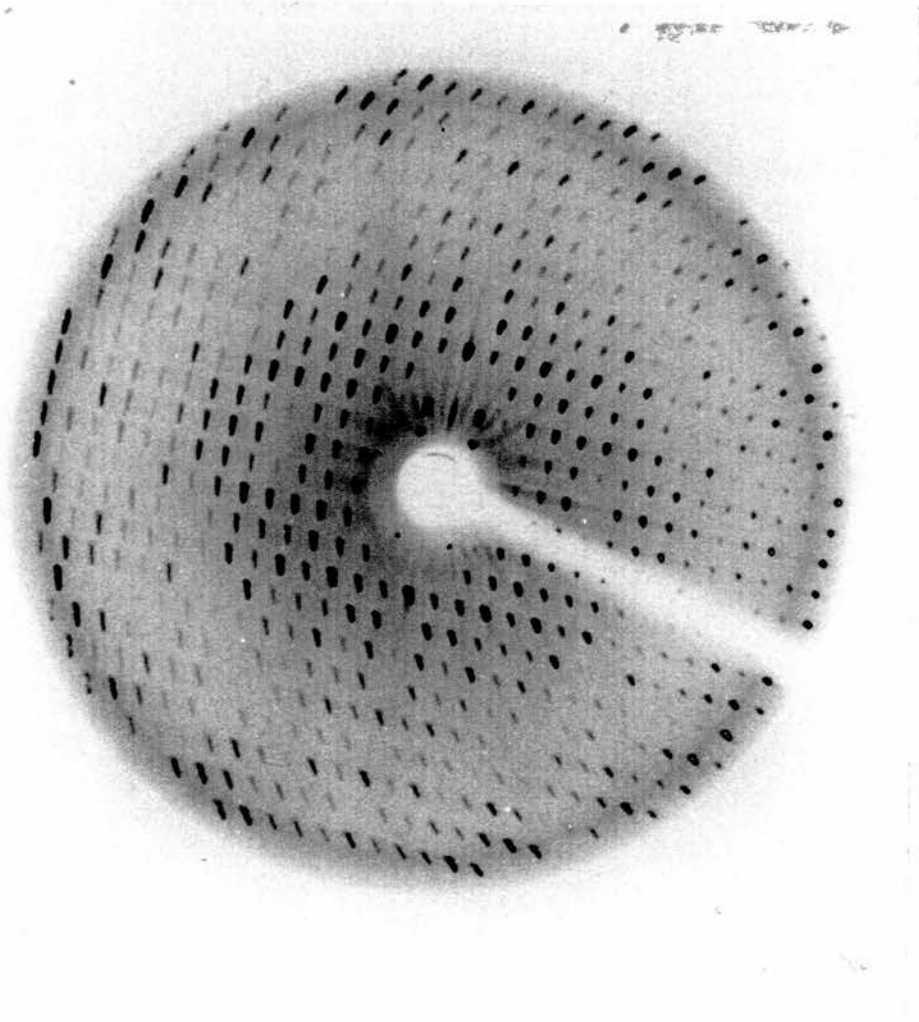


Figure 3–4: 10° precession photograph of hk0 projection collected on form B crystals

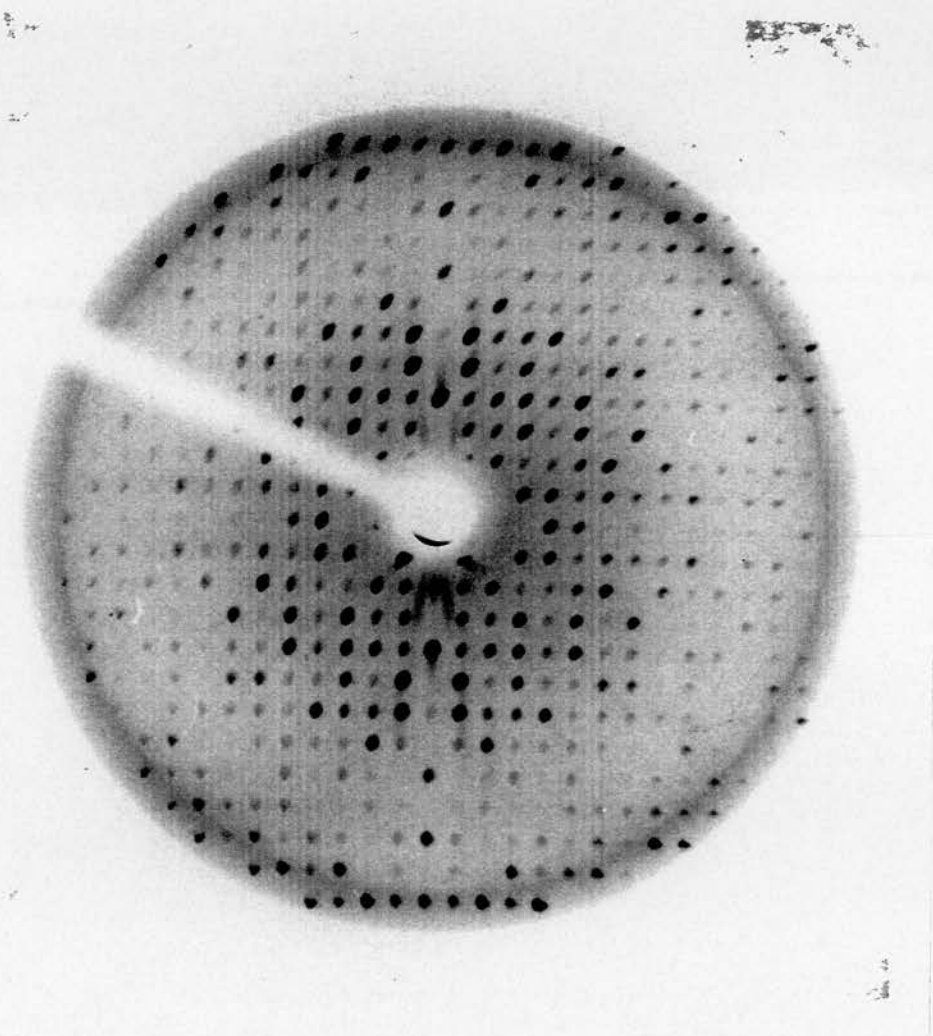


Figure 3–5: 10° precession photograph of 0kl projection from form B crystals.

Resolution (Å)	NATX				CMN			
	All Reflections		$h+l$ odd		All Reflections		$h+l$ odd	
	$< \frac{F}{\sigma F} >$	No.refs.	$< \frac{F}{\sigma F} >$	No.refs.	$< \frac{F}{\sigma F} >$	No.refs.	$< \frac{F}{\sigma F} >$	No.refs.
20.0 - 6.36	57.74	3160	1.74	1585	50.19	2348	0.67	1183
6.37 - 4.49	51.89	6254	1.63	3137	45.22	4285	0.72	2153
4.49 - 3.67	47.44	7505	1.67	3751	37.66	5439	0.79	2725
3.67 - 3.18	33.81	6906	1.61	3473	24.91	6250	0.91	3154
3.18 - 2.84	23.87	5940	1.40	2981	16.20	6763	0.94	3426
2.84 - 2.59	17.26	6201	1.39	3117	10.06	4436	0.98	2324
2.59 - 2.40	13.52	6250	1.25	3151				
2.40 - 2.25	10.48	5449	1.27	2751				
2.25 - 2.12	8.7	1678	1.20	853				

Table 3–2: Table showing the intensity of $h+l$ odd reflections of the large cell as a function of resolution. In the NATX data there are no reflections with $h+l$ odd and with $F/\sigma F > 6$, for the CMN data there are 13 reflections which fit this condition.

slightly with increased resolution, whereas the rest of the intensities for the rest of the data fall off rapidly with increased resolution.

3.2.3 Native data

A native data set was collected on a Xentronics detector on a Rigaku rotating anode generator (40 kV, 60 mV). Data were collected on the form B crystals as 0.25° images. The high resolution data were collected first (these data are more susceptible to radiation damage and therefore must be collected within a short time of the crystals being exposed to X-rays) by swinging the detector θ=-20° to one side of the beamstop and scanning through 180°. Completeness of the high resolution data was ensured by collecting 'cusp' data, that is moving the crystal so that it is rotated about an axis perpendicular to its initial rotation axis. The detector was then moved to θ -10° and another 180° scan plus cusp scan were made to collect the low resolution reflections. Images were processed using XDS (Kabsch 1988b). The output intensity data from XDS were scaled in two different ways and the data output by these programs compared using SCALEIT. Details of the data quality for the native form B crystals as scaled using ROTAVATA and AGROVATA (NATRA) are given in table 3–4. Statistics after scaling native data for form B crystals using XSCALE (NATX) are shown in table 3–3.

SCALEIT was used to analyze the differences between the two native data sets,

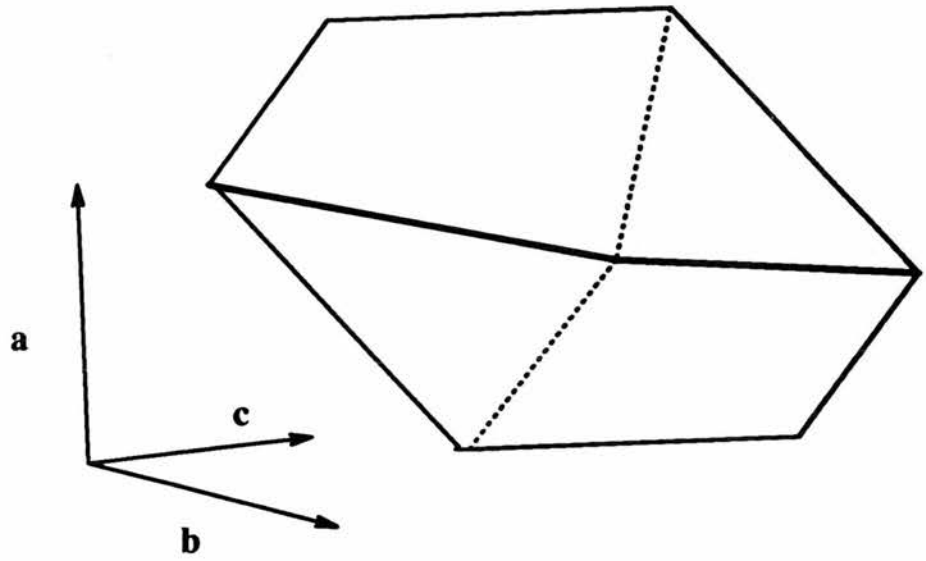


Figure 3–6: Diagram showing the relationship of the unit cell axes to the morphology of the form B crystals.

Resolution (Å)	N_{obs}		Total	R_{sym} %	Completeness %
	Observed	Unique	Unique		
5.60	5351	1517	1519	3.2	99.9
4.60	8554	2471	2482	3.3	99.6
3.80	13358	4114	4296	3.6	95.8
3.00	17899	9082	10032	3.9	90.5
2.60	11775	9018	10588	4.7	85.2
2.20	15271	13153	19604	6.5	67.1
Total	77487	41213	48521	3.5	84.9

Table 3–3: Crystallographic statistics and completeness as a function of resolution for the native data collected on DlADH form B crystals. Data were collected on a Xentronics detector (40kV, 60mA), processed using XDS and scaled using XSCALE.

Resolution (Å)	N_{obs}		R_{sym} %	Completeness	Redundancy
	Observed	Unique			
6.96	5125	1662	2.4	98.9	3.1
4.92	10291	2949	2.8	99.6	3.5
4.02	12193	3701	3.6	98.4	3.3
3.48	12149	4158	5.7	94.2	2.9
2.84	14302	9435	6.5	90.1	1.6
2.46	12933	10229	8.7	83.5	1.3
2.20	9954	8839	9.7	63.6	1.1
Total	76947	40973	2.9	82.5	1.9

Table 3–4: Crystallographic statistics as a function of resolution for the native data set of DlADH form B crystals. Data were collected on a Xentronics detector (40 kV, 60 mA), processed using XDS and scaled using ROTAVATA and AGROVATA

which had been scaled using different methods. Analysis of the native data showed that there were several large differences between the two data sets which can only be due to the different scaling techniques employed. But the weighted R-factor between the two data sets is 2.8% for 20029 reflections, indicating that there is no significant change between the data sets. However, the normal probability analysis (see section 5.1.4) was also used to compare the two data sets and this shows (see Table 3–5) that the standard deviations for the data scaled using XSCALE is an overestimate of the true standard deviations, and this overestimation is worse at high resolution. ROTAVATA and AGROVATA constrain standard deviations so that they have a normal distribution which has a mean of zero and a standard deviation of 1.0. If the standard deviations of the XSCALE data set are correctly estimated then the gradient of the normal probability plot would be 1.0. Since this gradient is less than one (see Table 3–5) then the standard deviations of the XSCALE data must be overestimated. There are two possible causes of this overestimation,

- The XSCALE program itself
- or an underestimation when converting XSCALE output intensity data to LCF format which used $\sigma F = \frac{\sigma I}{2F}$ to calculate σF.

Resolution (Å)	Gradient	Intercept
12.6	0.227	0.019
7.3	0.573	0.155
5.7	0.607	0.029
4.8	0.630	0.029
4.2	0.623	0.008
3.8	0.588	-0.016
3.5	0.549	-0.028
3.3	0.448	-0.039
3.1	0.306	-0.030
2.9	0.290	-0.038
2.6	0.276	-0.044
2.3	0.274	-0.042
2.2	0.285	-0.037

Table 3–5: Normal probability distribution as a result of comparing the native data scaled in different ways: XSCALE data compared to data scaled with ROTAVATA/AGROVATA. This analysis is dependent on the accurate estimation of standard deviations. The standard deviations for the data set should have a normal distribution with a mean of 0.0 and a standard deviation of 1.0. ROTAVATA and AGROVATA manipulate the data so this is true, but this analysis shows that the standard deviations for the XSCALE data are less than 1.0 and they are therefore underestimated.

Data Set	XSCALE data	ROTAVATA/AGROVATA data
Max. Difference	7768	6843
$< D_{iso} >$	672.3	734.6
Resolution (Å)	20-2.5	20-2.5
Total gradient	5.27	7.78
R_{iso}	16.9%	18.1 %
Scale factor	1.009	1.082
No. refs.	10473	9902

Table 3–6: Analysis of derivative data using native data that has been scaled using a) XSCALE b) ROTAVATA/AGROVATA.

The data were further analysed by using each native data set in turn (XSCALE and ROTAVATA/AGROVATA versions) to assess the quality of data collected on a heavy atom soaked crystal. Preliminary analysis of the native (both XSCALED and ROTAVATA/AGROVATA scaled) with the PTC data, was carried out in SCALEIT. The results of the analysis are summarized in table 3–6. The overall R_{iso} of the data sets show a slight change between the data sets. The probability analysis seems to indicate a more significant change between the data sets, but as can be seen above this is a function of the standard deviations of the data sets. It is not known how the standard deviations in XSCALE are calculated or standardized. There are some differences when scaling the PTC data to the differently scaled native data sets and these differences are apparent in the difference Patterson maps calculated for both sets of difference data (see Figure 3–7).

3.2.4 Data on soaked crystals

All data collected on heavy atom soaked crystals, were collected on a Xentronics detector. Data were processed with XDS and internally scaled with ROTAVATA and AGROVATA, LCF versions, before being converted to MTZ format. A summary of the data collection statistics and further analysis of the soaked crystal data used for isomorphous replacement are shown in chapter 5.

Platinum chloride data

Three data sets for K_2PtCl_4 soaked crystals were collected. Crystals were prepared by soaking in:

- 0.1mM $PtCl_4$ for 20 hours, in buffer with DTT (Data set referred to as PTC)
- 0.5mM $PtCl_4$ for 20 hours, in buffer with DTT (Data set PTC2)
- 1.0mM $PtCl_4$ for 20 hours, in buffer with DTT (Data set PTC3)

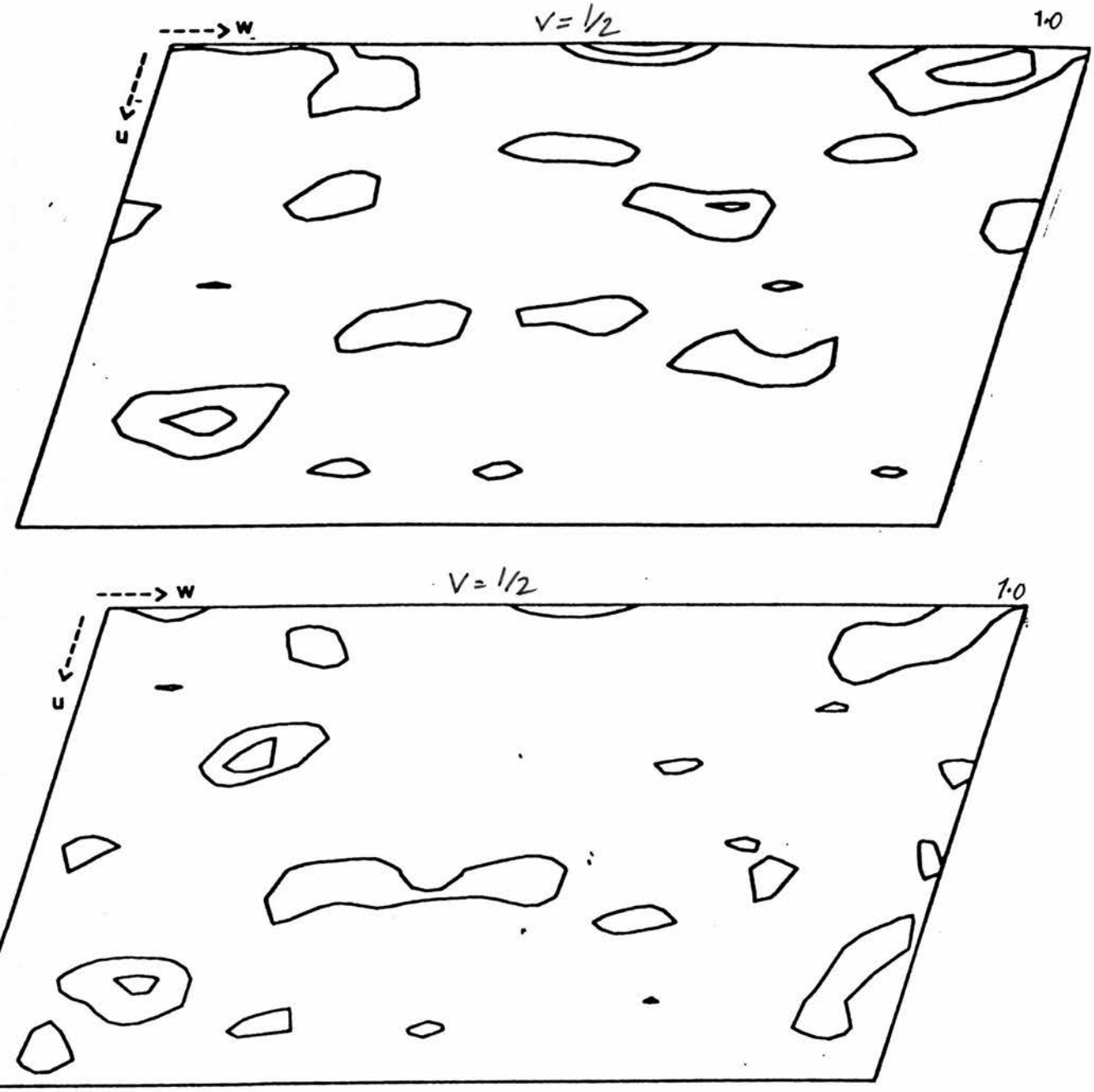


Figure 3–7: Difference Patterson map for (i) PTC data with native data from XSCALE and (ii) PTC data with native scaled with ROTAVATA/AGROVATA.

2-chloromercuri-4-nitrophenol data

Two data sets were collected on crystals that had been soaked in 2-chloromercuri-4-nitrophenol. One crystal was soaked in a buffer containing DTT, and later analysis showed that there was no substitution of heavy atoms (data set referred to as CMN). The mercury should bind to the free sulphydryls in the DlADH enzyme, however binding at these sites may be prevented by the presence of DTT in the soaking buffer. The second soak was prepared by removing DTT prior to adding heavy atom (the data set collected on this crystal is referred to as GCMN).

Mercury Chloride

A single data set was collected on a crystal prepared by soaking for 20 hrs in 3 mM $HgCl_2$. The crystal had been back soaked to remove DTT, prior to soaking in heavy atom solution (data referred to as GHG).

Tetrammineplatinum (II) chloride

A single data set was collected on a form B crystal soaked in 5 mM Tetrammineplatinum (II) chloride (data set referred to as TAPC).

3.2.5 Image plate

Data were collected on the MarResearch image plate at PX station 9.5, at the SRS Daresbury, U.K. (λ=1.07 Å). The crystal was rotated about **b***, and data collection started with **a** parallel to the incoming X-ray beam. The crystal to IP distance was 170 mm and the crystal diffracted to approximately 1.9 Å. The data were collected as 1.5° oscillation pictures over a range of 180°. The maximum oscillation range possible ($\Delta\phi$), is a function of the maximum resolution and the longest cell dimension. Hence,

$$\Delta\phi = \frac{p*}{dm*} - \Delta \qquad (3.1)$$

where $p*$ is the relevant reciprocal lattice spacing (that is the reciprocal lattice spacing parallel with the incident beam) and $dm*$ is the maximum resolution in reciprocal space. Δ is the angular separation of adjacent spots, such that:

$$\Delta_{2\theta} = \frac{p*}{cos\theta} \tag{3.2}$$

With the shortest reciprocal lattice vector parallel to the incoming X-ray beam, the maximum oscillation is 1.1° for data to 2Å resolution. However, in view of the limited time available the oscillation range was increased to 1.5° in order to reduce the total data collection time, but this will cause some overlapping of high resolution reflections. In retrospect, it would have been better to rotate the crystal about **c**. Before the onset of data collection two still pictures were taken at ϕ=90° and 180°, these are used in refinement of the orientation matrix and cell. Data were processed by MOSFLM (original suite originates from MOSCO system developed by Nyborg and Wonacott (1977) extensive modifications by A.J. Wonacott, P. Brick and A.G.W. Leslie). The processing package has the following order:

- Determination and refinement of the crystal parameters and orientation matrix, STILLS/IMSTILLS, REFIX, IDXREF and POSTCHK
- Generation of a list of predicted reflections, OSCGEN and reflection integration MOSFLM
- Data reduction, ABSCALE, ROTAVATA and AGROVATA.

IMSTILLS was used to create a file containing a list of the strong reflections from the first oscillation image. REFIX (Kabsch, 1988a and modified by Howard Terry) was used for initial autoindexing; approximate cell dimensions were given. The oscillation picture indexed with a positional residual of 0.08 mm and cell dimensions: a = 81.48, b = 55.45, c = 111.16 Å and β=94.72°. IDXREF was used to refine these initial parameters. IDXREF uses reflections recorded on two still images (for a monoclinic spacegroup, two still images separated in ϕ by 90° are used because of the low symmetry of the spacegroup). IDXREF failed in refinement. There are two possible reasons for this error:

- Crystal slippage between recording the still images and the oscillation image. If the crystal continues to slip it is difficult to obtain an accurate orientation matrix.

- Movement of the beam will affect the partiality of some reflections, since it is the partial reflections that are used to refine the orientation matrix, refinement will fail.

To overcome errors due to crystal slippage, an alternative procedure for refining the cell and the orientation matrix was tried. The first 30 images of the data were processed using the matrix from REFIX, the positional residuals from this run, between the predicted and observed reflection positions, were greater than 2 mm. REFIX has been found to give errors of greater than 0.1° in orientation angles or of 1% in unit cell dimensions. With this degree of error it is impossible to process IP data, since this system requires a greater degree of accuracy in cell parameters and crystal orientation to achieve the same positional accuracy as attained with film (due to the large detector surface). POSTCHK was used unsuccessfully to refine the cell parameters and orientation matrix on the data processed with the REFIX matrix. Since POSTCHK has a small radius of convergence which can be improved by several rounds of OSCGEN, MOSFLM and POSTCHK. Cycling through these programs was tried, but failed to refine the crystal orientation probably because the crystal had continued to slip during data collection and the degree of slipping was too great for POSTCHK to recover.

3.3 Discussion

Data collection has become a very much easier process in the last few years. Detectors are better and data collection is almost fully automated. However, there are many problems which can and do arise due to:

- Machine failure

- Poor data collection strategy

- Poor crystals

Good data is invaluable, care and attention taken during data collection can save time in later stages and lead to good interpretable maps. Good data redundancy leads to increased accuracy of data, this is especially important with native data. It is also important that the same data collection strategy is adopted for derivative and native data collection, so that the data are comparable.

The initial studies carried out on the form A, DlADH crystals show how difficult data processing can be if the crystals diffract weakly. Data collection becomes a much easier proposition when good crystals are available. Form B crystals, diffract X-rays well to high resolution and have posed few problems as far as data collection is concerned. Initial characterization of these crystals revealed a monoclinic unit cell with dimensions a = 81.2, b = 55.8, c = 109.7 Å, and β = 94.5° the cell is pseudo-centred, such that $h + l$ odd reflections are missing. The centring results in a smaller cell, a = 70.6, b = 55.8, c = 65.7 Å, and β = 106.9°, being valid. Although all data were processed in the large cell, for molecular replacement and isomorphous replacement studies the data were reindexed to a smaller cell. Analysis of the $h + l$ = odd reflections show that on average these reflections have intensities below the noise level, and therefore the small cell is valid in all instances. However, the autoindexing routines used to index the data routinely index the large cell, and since these routines use strong reflections it indicates that there are strong reflections which have indices $h + l$ odd.

A good native data set is invaluable and is one of the most important requirements for a successful X-ray structure determination. The native data collected on DlADH form B crystals, have a low R_{sym} and are complete to 2.5 Å. However, molecular replacement studies using this data gave inconsistent results (see Chapter 4), which were not comparable to results obtained for the CMN data. Self rotation function studies indicated that the solution for the CMN data was correct. As a result of the discrepancy in the molecular replacement study the native data were analysed and compared to the CMN data. The native data

set was missing 32 reflections between 4 Å and 12 Å resolution. Completeness of data is essential for molecular replacement studies and it is possible that these missing reflections result in very different solutions being obtained. The effect of these missing reflections is also discussed in chapter 5, where both the native and the CMN data sets are used to calculate difference Fourier maps for derivative data.

The effect of different scaling methods on the native data have also been examined. Theoretically, the local type scaling in XSCALE should give better internal scaling for data sets. However, the disadvantage of this method is the poor estimation of the standard deviations, as revealed by the normal probability distribution analysis. Standard deviations are used in weighting reflections, overestimation of standard deviations will introduce weak and badly measured reflections into certain calculations, for example, in refinement programs. An overestimation in the standard deviations in the initial stages of data processing will need to be accounted for in later processing.

Chapter 4

Molecular Replacement

4.1 Introduction

This chapter describes the molecular replacement study carried out for the structural determination of alcohol dehydrogenase from *Drosophila*. The 3-dimensional structure of 3α, 20β-hydroxysteroid dehydrogenase (HSD) from *Streptomyces hydrogenans*, was used as the search model. This search model was chosen since it is the only known structure of a 'short chain' dehydrogenase (Ghosh *et al.*, 1991). The sequence identity between the HSD and DlADH is only 23%, which is low for an attempt at molecular replacement. Usually a high level of sequence identity between the search model and the unknown is needed before a molecular replacement study is viable, for example a sequence identity of 30% (but preferably 50%) indicates a conserved structural motif.

4.1.1 Principles of molecular replacement

Molecular replacement can be used:

- To locate the noncrystallographic symmetry within a crystal system.
- To obtain initial phases for the structural determination of an **unknown** protein molecule by using the **known** structure of an homologous molecule.

Often when proteins crystallize, they crystallize with one or more subunits in the asymmetric unit. These molecules are often related by noncrystallographic symmetry, that is symmetry which is local to the asymmetric unit. This contrasts to crystallographic symmetry which extends throughout the entire crystal lattice. Noncrystallographic symmetry is often useful because it gives a greater redundancy of information for the asymmetric unit. The first step in any molecular replacement study is to establish the nature of any noncrystallographic symmetry (Rossmann and Blow, 1962). Noncrystallographic symmetry is found by rotating the Patterson function (from the crystal system under examination)

on its self, the noncrystallographic symmetry lies where the overlap is a maximum.

Many molecules can crystallize in several different crystal forms. If the structure has been determined in one crystal form then it must be possible to use the information from that structure to determine the coordinates of the molecule in a different crystal form (assuming that the only differences between the two crystal forms are due to different crystal contacts and that these changes are not gross). As new structures emerge, it has been seen that there exist families of structurally similar proteins, and in general each new structure is similar either in its entirety or has fragments of its structure that are similar to existing structures. If we can determine the relative orientation and position of a known structure (or fragments of structures) in the unknown's unit cell, then it is possible to include phase information derived from the suitably positioned known structures, into the unknown structure determination. If the similarity between the known and unknown structure is high, then the phase information derived from the known structure may be sufficient to solve the unknown structure directly.

To obtain the phases for an unknown molecule using molecular replacement techniques the known structure needs to be placed in the unknown's unit cell at the position of the unknown structure. Computationally the most efficient way of doing this is by:

- Orientating the known molecule in the same way as the unknown structure.
- Finding the position of the known structure in the unit cell of the unknown, with respect to the elements of crystal symmetry.

Similar structures must have similarities in their diffraction patterns. The diffraction pattern is a convolution of two sets of information: diffraction information from the crystal lattice and diffraction information from the contents of the asymmetric unit. The diffraction from the asymmetric unit is called the molecular transform and is a continuous function. The diffraction pattern from a

crystal is the molecular transform that has been sampled at discrete points, the form of this 'sampling' is dependent on the nature of the crystal lattice. When comparing diffraction information from two crystal systems the Patterson function is a convenient function to use: it can be calculated from the recorded intensities for the unknown and easily calculated using fast Fourier transform (FFT) methods for the model. A Patterson function contains no phase information as no origin in the unit cell is implied, only the relative positions of the atoms. The Patterson function is the convolution of the electron density of the contents of the unit cell with its centrosymmetric image (Patterson, 1934).

$$P_{uvw} = \int_0^1 \int_0^1 \int_0^1 \rho(\mathbf{x}) * \rho(\mathbf{x}+\mathbf{u}) dx.dy.dz \tag{4.1}$$

Where P_{uvw} is the Patterson function; $\rho(\mathbf{x})$ is the electron density at a point x, y, z in the unit cell; and $\rho(\mathbf{x}+\mathbf{u})$ is the electron density at a point $x+u, y+v, z+w$ in the unit cell. The integral is taken over the whole unit cell.

Since

$$\rho(\mathbf{x}) = \frac{1}{V} \sum_{\mathbf{h}=-\infty}^{+\infty} \mathbf{F_h} exp(-2\pi i \mathbf{h}.\mathbf{x}) \tag{4.2}$$

and

$$\rho(\mathbf{x}+\mathbf{u}) = \frac{1}{V} \sum_{\mathbf{h'}=-\infty}^{+\infty} \mathbf{F_{h'}} exp(-2\pi i \mathbf{h'}.(\mathbf{x}+\mathbf{u})) \tag{4.3}$$

where V is the volume of the unit cell, and where $\mathbf{x}$ and $\mathbf{h}$ are the vectors (x, y, z) and (h, k, l). The summation over $\mathbf{h}$ represents the summations over h, k, l. The resulting Patterson function is shown below. This integral vanishes unless $\mathbf{h} = -\mathbf{h'}$

$$P(u, v, w) = \frac{1}{V} \sum\sum \sum_{all\ hkl} |F_{\mathbf{h}}||F_{-\mathbf{h}}| exp(2\pi i \mathbf{h}.\mathbf{u}) \tag{4.4}$$

and, since

$$F_{\mathbf{h}} F_{-\mathbf{h}} = |F^2| e^{i\phi} e^{-i\phi} = |F^2| \tag{4.5}$$

where ϕ is the phase with respect to the origin of the unit cell and this phase information is lost from the Patterson function. The information left in the

equation is the relative positions of the atoms.

$$P(u,v,w) = \frac{1}{V}\sum\sum\sum_{all\ h,k,l} |F|^2 exp(2\pi i \mathbf{h}.\mathbf{u}) \tag{4.6}$$

Friedels law states that, $|F|^2(hkl) = |F|^2(-h,-k,-l)$. Therefore, $e^{i\phi}$=$e^{-i\phi}$ since $e^{i\phi} = cos\phi + isin\phi$ and because $cos\phi = cos(-\phi)$ and $sin\phi = -sin(-\phi)$, the sine terms cancel out and the Patterson function can be written:

$$P(u,v,w) = \frac{1}{V}\sum_{h\geq 0}\sum\sum_{all\ k,l} |F|^2 cos(2\pi \mathbf{h}.\mathbf{u}) \tag{4.7}$$

The Patterson function represents a vector map of the electron density in the unit cell, as such it is composed of 'self vectors' and 'cross vectors'. 'Self ' or 'intramolecular' vectors are vectors between atoms in the same molecule. 'Cross' or 'intermolecular' vectors relate atoms in different molecules. Examination of the 'intramolecular' vectors gives information on the symmetry of the molecule while the 'intermolecular' vectors provide information about the positions of the molecules with respect to one another.

4.1.2 The rotation function

The rotation function is used to correlate a spherical volume of a given Patterson function with itself, or another Patterson from a similar molecule, with one Patterson function being rotated upon the other and the correlation function evaluated for each rotation step. In this chapter we meet three conventions for rotational systems:

Spherical polar rotation: describes a system where the self rotation axis is tilted at $\omega°$ to the Z axis and at $\phi°$ to the X axis. κ is the degree of rotation about this defined axis; for a 2-fold rotation κ=180° (as defined in POLARRFN).

Eulerian angles: (Rossmann and Blow, 1962) describe a set of rotations with respect to an orthogonal system; a rotation of α about Z, followed by a rotation β about the new Y and a rotation γ about the new Z. All rotations are anti-clockwise looking along the axis towards the origin.

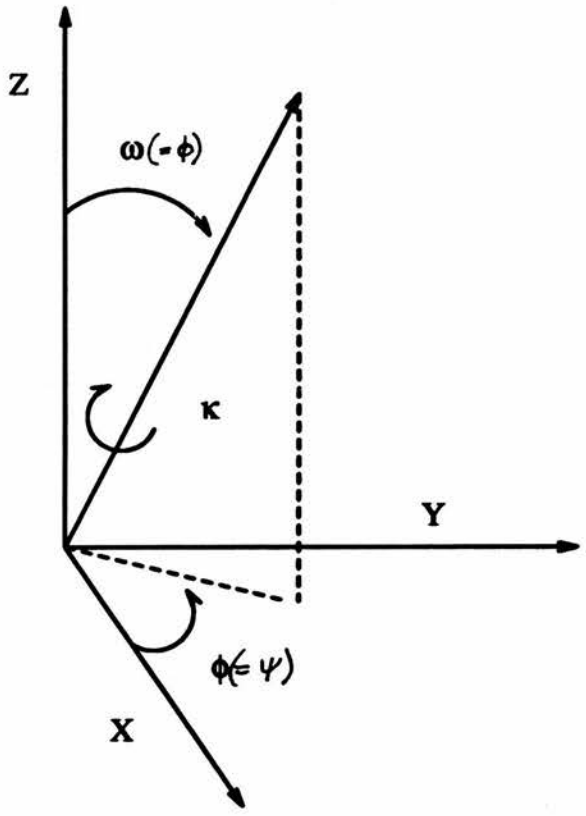


Figure 4–1: Spherical polar coordinate system as defined in **POLARRFN** (as defined in MERLOT).

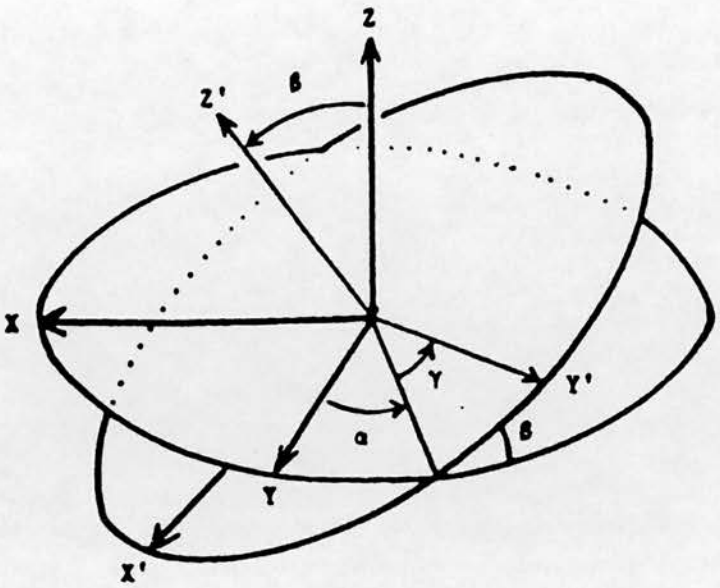


Figure 4–2: Euler angle coordinate system. Where α is the rotation about Z; β is the rotation about the new Y axis; and γ is the rotation about the new direction Z' (Blow, 1985).

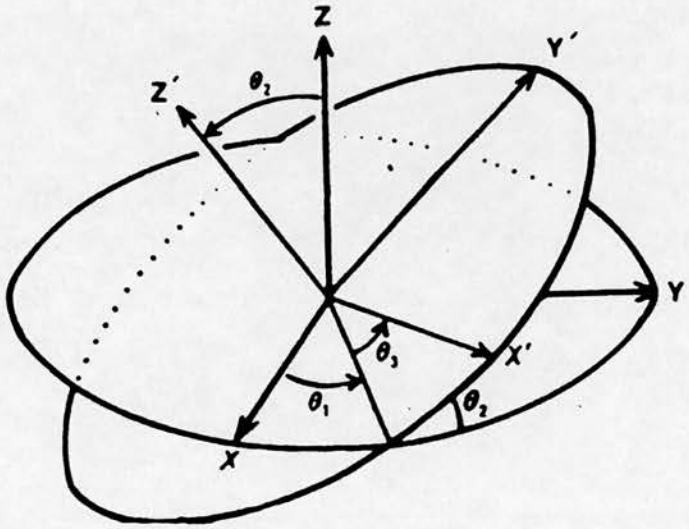


Figure 4–3: Lattman psuedo Eulerian angles. θ_1 is the rotation about Z; θ_2 is the rotation about the new X axis; and θ_3 is the rotation about the new Z position (Lattman, 1985).

Lattman's pseudo Eulerian angles: (Lattman, 1985) these angles describe a rotation θ_1 about Z (with a positive rotation being anti-clockwise about Z when viewed looking towards the origin), θ_2 about new X axis and θ_3 about new Z axis. Crystal symmetry is expressed less fully in polar angles, therefore they are often used to compute self rotation functions while Euler angles are used in cross rotation studies.

A general form of the rotation function, as presented by Rossmann and Blow (1962),

$$R(\mathbf{C}) = \int_{-\infty}^{+\infty} P_1(\mathbf{x}) U(\mathbf{x}) P_2'(\mathbf{Cx}) dV \tag{4.8}$$

Where P_1 and P_2 are Patterson functions for the model and unknown respectively; U is a shape function and **C** is a rotation matrix. When determining the relative orientation of two molecules, it is the rotation matrix used to rotate one Patterson function onto the other, that is being sought. For convenience the rotation function is generally computed in reciprocal space:

$$R(\mathbf{C}) = \sum_p \sum_h F_2(\mathbf{p}) F_1(\mathbf{h}) G(\mathbf{h} + \mathbf{h}') \tag{4.9}$$

F_1 and F_2 are the intensities that correspond to the reciprocal lattice vectors P_1 and P_2. The non-integral lattice vector h' is given by $\tilde{\mathbf{C}}\mathbf{p}$ where $\tilde{\mathbf{C}}$ is the transpose of **C** and **p** is the reciprocal lattice vector. G is the Fourier transform of the shape function U, and is an interference function. Rossmann and Blow (1962) used a 'product function' to evaluate the rotation function, others have used a 'sum function' or 'minimum function'. Reciprocal space programs are quicker than the real space equivalents, but the real space methods seem to be slightly more accurate (M. Buehner and H. J. Hecht, 1985).

The computation of the Rossmann and Blow rotation function has been modified so that it is computationally swifter. Lattman's rotation function is written,

$$R(\mathbf{C}) = \sum_p F_M(\tilde{\mathbf{C}}\mathbf{p}) F_2(\mathbf{p}) \tag{4.10}$$

Where R(**C**) is the rotation function; F_M is the Fourier transform for the isolated search molecule; ($\tilde{\mathbf{C}}\mathbf{p}$ is the rotation matrix multiplied reciprocal lattice vector **p**). F_2 is the intensity corresponding to the reciprocal lattice vectors of the unknown (P_2). The following assumptions are made: the sum over **h** is a convolution of F_1 with G, the interference function. This is equal to multiplying the Patterson with a shape function, U. U is no longer necessary when dealing with an isolated molecule and rotating P_1 instead of P_2.

Further work by Crowther (Crowther, 1972) showed that by expanding the Patterson density within a spherical volume in terms of spherical harmonics, (as opposed to working with Cartesian Fourier components), allows the use of Fourier transforms to evaluate the overlap function. This makes the calculation of the rotation function much quicker. But a degree of error is introduced into the solution due to the expansion in spherical harmonics. The suggested procedure for the rotation function when using the MERLOT suite of programs is to use CROSUM, a Crowther rotation function program to do a complete search over all of reciprocal space; then to use LATSUM, which uses the Lattman rotation function, to search a small region of space about the solution peak found using CROSUM. In this way an accurate solution to the rotation function should be found.

Another calculation for the rotation function has been implemented in the program X-PLOR, using a real space rotational search (Brünger, 1990; Huber, 1985). This rotation function uses pseudo-orthogonal Eulerian angles and is restricted to the asymmetric unit (Rao *et al.*, 1980). For each orientation the rotation function, RF(Ω), is computed.

$$RF(\Omega) = < P_{obs} P_{model}(\Omega) > \tag{4.11}$$

where P_{obs} is the interpolated Patterson computed from the observed intensities, P_{model} is the Patterson for the rotated model. X-PLOR uses a Patterson correlation (PC) refinement, applied to peaks found in the conventional rotation search, to find the best solution. The target function for PC refinement is a combination of the effective Patterson energy and an empirical energy term. The

Patterson energy term is proportional to the negative correlation coefficient between the squared amplitudes of the observed and calculated normalized structure factors. The empirical energy term conveys information about the geometry and non-bonded interactions of the molecule (Karplus and McCammon, 1983). During PC refinement the atomic coordinates, (or groups of atomic coordinates), in the search model are moved, such that the target function is minimized or the correlation coefficient is maximized. This method has been particularly successful when there are conformational differences between the search model and the unknown structure. The target energy is defined as:

$$E_{tot}(\mathbf{r}) = E_{PC}(\mathbf{r}) + E_i(\mathbf{r}) \tag{4.12}$$

where E_i is an empirical energy term which describes geometry and non-bonded interactions of the search model, and E_{PC} is an energy term that is proportional to the negative linear correlation coefficient PC($\mathbf{r}$,Ω) between the squared amplitudes of the normalized observed and model structure factors.

$$PC(\mathbf{r},\Omega) = \frac{[< |E_o|^2|E_m|^2 > - < |E_o|^2 >< |E_m|^2 >]}{[[< |E_o|^4 > - < |E_o|^2 >^2][< |E_m|^4 > - < |E_m|^2 >^2]]^{1/2}} \tag{4.13}$$

where

$$E_{pc} = W_{pc}[1 - PC(\mathbf{r},\Omega)] \tag{4.14}$$

Where E_o is the normalized observed structure factor and E_m is the normalized model structure factor. When using rigid groups of atoms the empirical energy term is turned off. PC($\mathbf{r}$,Ω) is proportional to the product function used in the conventional rotation function (Lattman, 1985) $< |F_o|^2|F_m|^2 >$.

Checks for the rotation solution

The results of the rotation function should be consistent between different resolution ranges and where possible between different data sets. Inconsistencies can be the result of:

- systematically missing intensity data
- gross structural differences between the search model and the unknown (these can be a function of resolution).

If there is any noncrystallographic symmetry, then the rotation function solution should move the search model to a position which generates this observed symmetry.

4.1.3 The translation function for positioning a correctly orientated molecule fragment

There are three main types of methods for positioning a correctly oriented molecule in the unknown cell with respect to the crystal symmetry elements (Beurskens *et al.* (1987) gives a more complete review of the literature on translation functions);

a) Patterson methods that correlate the Patterson function from the model system with the intermolecular vector set from the unknown structure. These methods are sensitive to accuracy of the rotation angles, systematically missing data and noncrystallographic symmetry which lies parallel to crystallographic symmetry axes.

b) R-factor searches that measure discrepancy between the observed intensities and the intensities generated from the model as it is positioned in the unit cell. This technique is sensitive to scaling. Also, the fact that it cannot be evaluated using Fourier techniques means that it is slow.

c) A correlation function, combines information from the crystal symmetry, the steric restrictions of the model and the diffraction data. The correlation function can be expressed as a Fourier series using FFT's, it is therefore computationally swift.

Patterson methods used for positioning a molecule in the asymmetric unit are based on the work of Crowther and Blow (1967). The translation function can be

calculated for the intermolecular vector set between two asymmetric units or over the whole unit cell.

$$T_1(\mathbf{t}) = \int P_O(\mathbf{u}) P_p(\mathbf{u}, \mathbf{t}) d\mathbf{u} \tag{4.15}$$

where the integration is over the whole unit cell

$$T_2(\mathbf{t}) = \int P_O(\mathbf{u}) [P_p(\mathbf{u}, \mathbf{t}) - \sum_{s=1}^{m} P_s(\mathbf{u})] d\mathbf{u} \tag{4.16}$$

T_2 is a full symmetry translation with intramolecular vector removed. The removal of intramolecular vectors from this calculation increases the signal to noise. $P_O(\mathbf{u})$ is the observed Patterson function at a position $\mathbf{u}$. $P_p(\mathbf{u}, \mathbf{t})$ is the calculated Patterson function for the orientated and positioned search molecule (and its symmetry related mates), this term includes intramolecular vectors. $P_s(\mathbf{u})$ is the self vector set for the molecular fragment, which is independent of $\mathbf{t}$ and can therefore be subtracted. Langs (1985) has proposed a translation function formalism in which the structure factor terms for the model structure (intensity and phase) are replaced with phase information only. He also proposes a method for reducing the structure-dependant spurious maxima (as long as the search model is conformationally true). This method does not work well when trying to position a fragment of the search model ($< \frac{1}{4}$ of the unknown molecule) in the unknown cell. This formalism is utilized in MERLOT.

The correlation method for determining the translation function solution uses a correlation coefficient, $TO(\mathbf{t})$, combined with a packing function or overlap function, $O(\mathbf{t})$. Therefore, $T(\mathbf{t})$ pools information from real and reciprocal space (Harada *et al.*, 1981).

$$TO(\mathbf{t}) = \sum |F_o|^2 |F_c|^2 \sum |F_o|^4 \tag{4.17}$$

$$O(\mathbf{t}) = \sum \frac{|F_c^2|}{n} \sum |F_m|^2 \tag{4.18}$$

$$T(\mathbf{t}) = TO(\mathbf{t}) / O(\mathbf{t}) \tag{4.19}$$

Again, F_o and F_c are the observed and calculated structure factors, respectively. n is the number of symmetry operations and F_m is the contribution of the

structure factor from one molecule. This resulting function is a scalar product and as such is less sensitive to errors in the orientation of the molecular fragment. It is possible to compute $TO(\mathbf{t})$ using an FFT. It is essential to use E_o and E_c to compute $TO(\mathbf{t})$

The $T(\mathbf{t})$ correlation function has been modified so that it is insensitive to scaling and is used in BRUTE (Fujinaga and Read, 1987).

$$C = \sum(|F_o^2| - \overline{|F_o|^2})(|F_c| - \overline{|F_c|^2})[\sum(|F_o|^2 - \overline{|F_o|^2})^2 \sum(|F_c|^2 - \overline{|F_c|^2})^2]^{-}1/2 \quad (4.20)$$

Where C is the correlation factor and it is best calculated using the normalised structure factors (or by using a narrow resolution range of data which is a crude method of sharpening the data). The value of C is dependent on the fraction of the crystal content that the model represents and on the degree of homology between the model and the unknown structure. The program positions the search molecule in the unit cell, then generates the symmetry equivalents. The structure factors are calculated using the molecular scattering factors (Lipson and Cochran, 1957). The search molecule is moved through the unit cell grid point by grid point. At each grid point the structure factors are calculated and the correlation function is computed. The time taken to compute this function is dependent on the number of symmetry operations, the number of reflections and the number of grid points. The correlation coefficient is a flat function which requires that a very fine grid is used so that the solution peak is not missed, therefore these are long calculations. BRUTE uses two methods to 'refine' the orientation which in turn optimise the translation function solution. One method shifts the orientation slightly and then carries out the translation function with each new orientation, however this can be computationally very expensive. The alternative is to place the search model in a P1 cell and then rotate the model until a maximum in the correlation coefficient is attained, before carrying out the translation function.

The translation function within X-PLOR, is also based on the linear correlation coefficient (Harada *et al.*, 1981) and is similar to BRUTE. However, it is statistically more efficient because it uses the normalized structure factors. The

X-PLOR translation function is given by:

$$TF = \frac{< |E_o|^2|E_c|^2 > - < |E_o|^2 >< |E_c|^2 >}{[< |E_o^4 > - < |E_o|^2 >^2][< |E_c^4 > - < |E_c|^2 >^2]^{1/2}} \tag{4.21}$$

Where E_o is the normalized structure factor for the unknown structure; E_c is the normalized structure factor calculated for the model. The program calculates the normalized structure factors and their symmetry mates, it stores them, then as the search model is moved through the unit cell appropriate phase shifts are applied to these calculated structure factors. A packing function (Hendrickson and Ward, 1976) is also incorporated into the translation search and evaluated so that the overlap of the positioned molecules is minimized.

4.1.4 Phased translation function

The phased translation function (PTF) involves computing the correlation between electron density of the unit cell, as derived from some previously determined phases, with the electron density from the translated model. The PTF be evaluated as a Fourier transform.

$$C(\mathbf{t}) = \frac{\int_V (\rho_P(x) - \overline{\rho_P})(\rho_M(x-t) - \overline{\rho_M})dx}{[\int_V (\rho_P(x) - \overline{\rho_P})^2 dx \int_V (\rho_M(x-t) - \overline{\rho_M})^2 dx]^{1/2}} \tag{4.22}$$

Where ρ_P is the electron density computed from prior phase information; ρ_M is the density of a single molecule with the correct orientation but an arbitrary position. The translation vector is $\mathbf{t}$ and $\overline{\rho}$ is the mean density in the unit cell with volume, V. Since the mean density is independent of the translation vector this equation can be simplified. Expanding the equation in terms of the complex structure factor, the phased translation function can be written:

$$C(t) = \frac{k}{V} \sum_{\mathbf{h}} |F_o(\mathbf{h})||F_M(\mathbf{h})| exp[i(\alpha_P - \alpha_M)] \tag{4.23}$$

where

$$k = \frac{V}{[\sum_{\mathbf{h}} (m_P |F_o(\mathbf{h})|)^2 \sum_{\mathbf{h}} |F_M(\mathbf{h})|^2]^{1/2}} \tag{4.24}$$

This technique is a very powerful method of determining the translation function, the information obtained from heavy atom derivatives can overcome

problems due to slight inaccuracies of the rotation function solution. However, in order to ensure that the handedness of the isomorphous phases are correct, the correlation function must be computed again using $(-\alpha_P - \alpha_M)$.

4.1.5 Refinement of molecular replacement solutions

There are several steps at which refinement can be carried out :

For the rotation function, refinement usually means repeating the rotation function over a small area of rotation space and using a fine grid search. PC refinement offers an alternative approach whereby the position of the search model can be adjusted while the agreement between the observed and calculated Pattersons are minimized.

Once the translation function has been solved, and the position of the search model in the unit cell determined, then more conventional refinement can be applied:

- Rigid body refinement moves the search model as the $|F_{obs} - F_{calc}|$ is minimized.

- A more forceful approach can use simulated annealing (Weis and Brünger, 1989) refinement where atomic positions are changed. Initially energy is put into the system, this allows the molecule to overcome any energy barriers which may have trapped it into a local energy minimum. In this way the refinement explores more conformational space.

- RMINIM in MERLOT, in contrast, applies small shifts to the position of the search model and then calculates the R-factor at each point.

4.1.6 Assessment of correctness

An assessment of the correctness of a preliminary molecular replacement solution is often a problem. But careful assessment of results at each stage of the

molecular replacement procedure should be carried out. The rotation and the translation functions are carried out over a number of different resolution ranges. Consistent solutions using data from different resolution ranges, seems to indicate a reliable solution. However, an inconsistent solution is not necessarily wrong, since data at different resolutions will provide contradictory information. For example, inclusion of high resolution data assumes a high degree of similarity between known model and unknown structure; and the inclusion of low resolution terms may introduce information about solvent regions of the crystal which will differ between the model system and the unknown.

A correct solution should refine using conventional rigid body refinement. The calculated R-factor, however, is not an indication of the quality of that solution since it is dependent on the completeness of the search model and it's similarity to the unknown.

The symmetry equivalents for the proposed solution should pack in the unit cell in a reasonable manner i.e. they should not penetrate each other. This packing can be done using a graphics package or CONTAC a subprogram within MERLOT.

The phases from the molecular replacement solution can be checked if there is any information available from heavy atom data. A difference Fourier for the heavy atom data, calculated using the phases from the molecular replacement solution, should agree with heavy atom positions located using the difference Patterson maps (see section 4.2.10).

The final check for a molecular replacement solution is to look at the map generated from the molecular replacement phases and see if it is interpretable.

4.2 Methods

The self rotation function calculation was carried out using POLARRFN (CCP4 program) and CROSUM (within MERLOT (Fitzgerald, 1988)) and X-PLOR (Brünger, 1990). The results were compared.

The cross rotation and translation function calculations were carried out using a poly-alanine model of the HSD, using both a dimer and a monomer, as the search model. The programs used were: CCP4 self rotation programs; the MERLOT suite; the molecular replacement routines within X-PLOR and BRUTE (Fujinaga and Read, 1987).

The cross rotation function was carried out using CROSUM in MERLOT, and the real space algorithm in X-PLOR. Often inaccuracies in the rotation function solution create problems in the translation function. Attempts to optimize the rotation used the Patterson correlation refinement in X-PLOR. Translation function calculations were carried out using TRNSUM in MERLOT, X-PLOR and BRUTE. When the translation calculations failed to find a solution, further studies with the rotation function were undertaken. Using the 'direct rotation search' in X-PLOR both rotation and translation solutions were found. This solution was refined but the resulting map was not good enough to give an unambiguous chain trace. However, the phases from the molecular replacement were good enough to locate heavy atom positions using a difference Fourier map. Initially, X-PLOR was run on an Evans and Sutherland ESV10, which was not efficient and routine use of the program only became possible when it was implemented on a Digital ALPHA T1.0 VMS system (Edinburgh University Computing Services, EUCS). BRUTE was implemented on the Connection Machine (CM-200), manufactured by Thinking Machines Corporation, where the calculation of the structure factors had been parallelized (P.D. Adams, personal communication).

4.2.1 Self rotation function

The self rotation function rotates the self vector set from an unknown structure onto itself. The aim in our case is to locate the elements of local symmetry within the molecule. Generally speaking, the self rotation function is used to locate local symmetry within the asymmetric unit of the unit cell. The self rotation was carried out using both the large cell and the small cell for the DADH crystals (see Chapter 3). In the large cell there are two dimers in the asymmetric unit, in the small cell only one dimer. The self rotation function was carried out in both cells to ensure that the reindexing did not affect the position of the molecular diad. Analysis showed that even when present, $h+l$ = odd reflections were very weak (that is less than $3\sigma_I$). The small unit cell is more convenient for molecular replacement calculations because there is only one dimer in the asymmetric unit. Initial attempts using MERLOT and the Crowther rotation therein, CROSUM, yielded unconvincing results to the self rotation function. This rotation is carried out using polar angles and a 'number field' type output is used for sections of constant κ. Peaks at ϕ=90.0°, ψ=72.0° and κ=180.0° had peak heights of 2-3σ, where σ is a measurement of the background signal (this depends on resolution range and integration radius). Peaks on this section generated additional peaks related by 90.0° in ψ because they lie perpendicular to the crystallographic two-fold axis. The self rotation was carried out on all of the data sets. The resulting outputs were significantly different. Further, the results of the self rotation function for different data sets were not consistent. Initially this was thought to be due to the degree of completeness of the data sets since: missing data affects the Patterson function calculated and hence the self rotation function. Table 5–2 shows the completeness of the four data sets used in this study, each is analysed as a function of resolution. The NATX and CMN data sets are reasonably complete to 4Å, and since both data sets were collected using a 180° scan plus cusp scan, it is assumed that any missing data is randomly distributed in reciprocal space. At low resolution (less than 6Å) the NATX data is the most complete. CMN and PTC data were collected with identical detector

settings, both are approximately 97% complete at low resolution which indicates that a section of the data were missed using this collection procedure.

The self rotation calculation was repeated using POLARRFN a fast rotation function program from CCP4. This program gave consistent results, as a function of data set; as a function of resolution and as a function of Patterson integration radius. The program produces sections of constant rotation angle κ and these sections are contoured at constant σ (where σ is a measure of the background signal) which are complemented by a peak listing. A noncrystallographic 2-fold axis would give rise to a peak on the κ=180° section (it is good practice to look at all κ sections regardless of the noncrystallographic symmetry expected).

The rotation function was calculated for different resolution shells and different radii of integration. The solution with the best signal to noise was found for data between 4-8 Å and with an integration radius of 23 Å.

The self rotation function peaks lie at ω=22.5, ϕ=114.2, κ=180.0 in the large cell and ω=22.5, ϕ=77.0, κ=180.0. Both solutions refer to the same diad (see Figure 4–4) and this solution was consistent for all data sets.

The self rotation was finally carried out using X-PLOR. The self rotation functions for the CMN, NATX and NAT data sets were calculated, using an integration radius of 23 Å and data between 8-4 Å resolution, all data sets gave the same molecular diad to within 3° and matched the POLARRFN solution.

The angle conventions in X-PLOR and POLARRFN are different. Both use spherical polar coordinate systems but ω in POLARRFN is equivalent to ϕ in X-PLOR. ϕ in the POLARRFN system is the rotation about Z, measured from **c**. ψ in the X-PLOR system is the rotation about Z but is measured from **a***. κ is the same in both systems. The X-PLOR results confirm the POLARRFN results.

The self rotation studies located a noncrystallographic diad axis at ω=22.5°, ϕ=76.5° and κ=180°. This solution was confirmed using POLARRFN and X-PLOR. The self rotation function within MERLOT did not give a consistent solution.

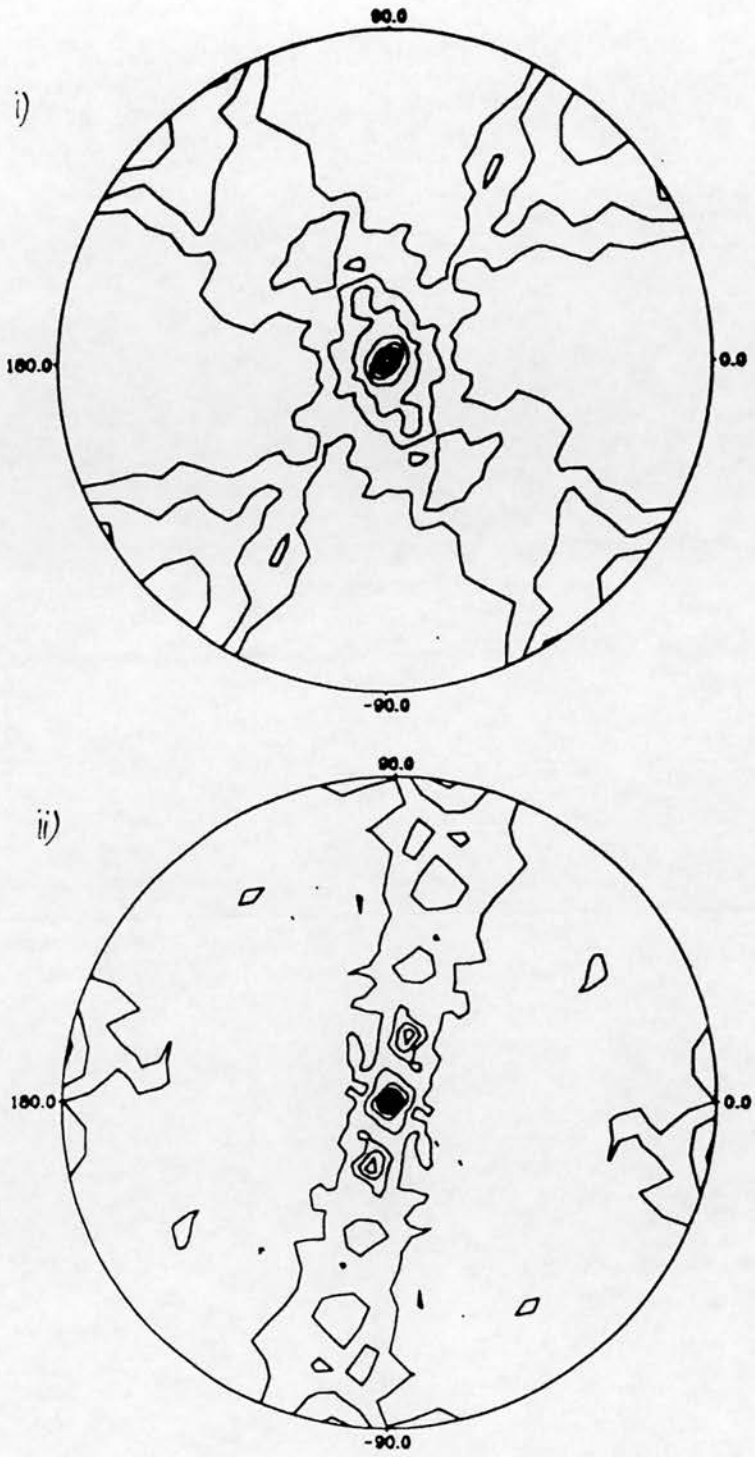


Figure 4–4: Self rotation functions for i) the large cell shows a peak at ω=22.5°, ϕ=114.2°, κ=180.0°; ii) the small cell shows a peak at ω=22.5°, ϕ=77.0°, κ=180.0°. The diad are related by a rotation of 37.2° about b (see Figure 3–3).

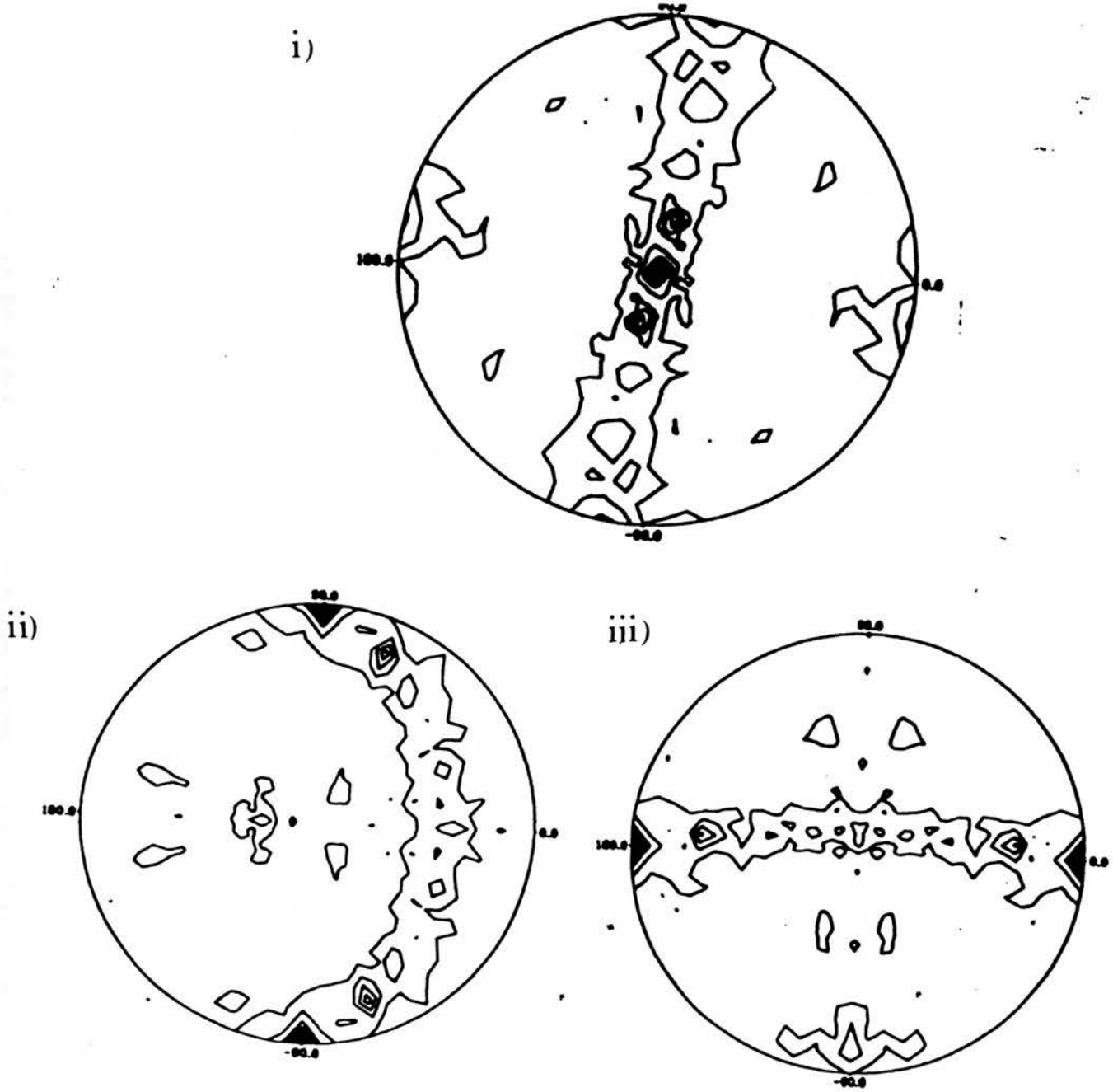


Figure 4–5: The self rotation function for 4-8 Å resolution in the small cell, with 23 Å radius for different orthogonalization conventions X, Y, Z are along: i) c,b*xc,b* ii) a, c*xa, c* iii) b, a*xb, a*, respectively

When a noncrystallographic diad lies parallel to a crystallographic axis then it will produce a large peak on the Harker section ($v = 1/2$), whose coordinates determine the position of the molecular centre in the xz plane. If there are N atoms in each subunit, then the origin peak of the Patterson for a monoclinic spacegroup is proportional to 4N because there are 4 subunits in the unit cell. The height of the peak due to the noncrystallographic symmetry peak is then proportional to 2N. Since the noncrystallographic diad in the case of DlADH, is tilted at 22.5° to the crystallographic diad then it is expected that this peak will be slightly offset and will provide only an approximation of the molecular centre (a simple geometrical relationship shows that a diad length 28 Å tilted to 22.5°, can lead to a displacement in the Harker section of more than 10 Å). The angle of the diad will also lead to a reduction in the height of the noncrystallographic symmetry peak. The Harker section for the native Patterson was examined and a peak was found which corresponds to a molecular centre of x = 0.20 and z = 0.27. This peak height is 10% of the origin peak height.

4.2.2 Cross rotation function

The cross rotation study aims to orientate a search model in the cell of the unknown structure by looking at the overlap of the Patterson functions of the known and the unknown molecules.

Choice of the search model

The degree of similarity between the search model and the unknown is an important factor in determining the success of the molecular replacement study. Similarity can be assessed by looking at sequence alignments, secondary structure predictions and known structural features of the molecules.

The search model used in this case was HSD. The active form of this enzyme is a tetramer and it crystallizes with one tetramer in the asymmetric unit (see Figure 1–2). Each subunit consists of a single domain composed of seven parallel β strands which form a doubly wound β sheet (Richardson, 1984). This sheet

Figure 4–6: 3α, 20β-hydroxysteroid dehydrogenase Q-axis dimer.

has three α helices on each side. The dinuceotide binding site is generated by the βA-αB-βB-αC-βC fold. The conserved GXXGXG sequence motif (common to most dehydrogenases) lies in the βA-αB turn. The structure continues αD-βD-αE-βE-αF-βF such that βD packs against βA. The αG lies next to αB and a long 32-residue loop links to βG and the final 16-residue carboxyl loop. In the tetramer, three two fold axes act between the monomers. These axes are labeled conventionally (Rossmann *et al.*, 1973) as P, Q and R. The αG helices, loop λG, the carboxyl terminal and the two amino termini associate about the P-axis. The Q-axis interaction (see Figure 4–6) is between the βD-αE turn, helix αE, βE-αE turn and helix αF. This gives a helix-helix and a helix-loop type packing. The association of βD-loopλD-αE of adjacent subunits enhances the structural stability. The R-axis interface has the fewest points of contact, the interactions are through the carboxy-terminal arm and the λE loop. The overall dimensions of the tetramer are 65 Å x 62 Å x 54 Å, along P, Q and R respectively. Each monomer has a single cofactor binding site and a single substrate site. The NAD^+ binds to the amino-terminal ends of the β-strands A, B and C and also to the midsections of βD and αE. The NAD^+ binds in an extended conformation. The substrate is thought to bind in a deep cleft and is in contact with residues from each subunit, which explains the occurrence of the tetramer as the active form.

The sequence alignment of DlADH and HSD is shown in figure 4–7. Sequence identity is low but other considerations encouraged us to continue with the molecular replacement study:

- The predicted occurrence of a Rossmann fold at the N-terminus region of the DADH (Thatcher and Sawyer, 1980; Villarroya, 1989).

- The highly conserved nature of the Rossmann fold which occur in proteins whose overall sequence similarity is poor. The Rossmann fold of some of the medium chain dehydrogenases have been superimposed to within 2 Å (Brändén and Tooze, 1991). Regions of highest similarity are the β1-αA-β2

motif including the loop regions, the major parts of helices B and D and the other β strands.

- The Rossmann fold is predicted to form a significant part of the DADH monomer, approximately half of the monomer.

To our knowledge, Rossmann folds are not used routinely to solve other structures by molecular replacement. Possible reasons for this are:

- The Rossmann fold comprises only a small fraction of the unknown structure, especially if the side-chains are not included in the calculation. e.g. for medium chain dehydrogenase the Rossmann fold comprises about one third of the whole subunit.

- The structural diversity for this fold is too great (r.m.s. 2 Å).

- The Rossmann fold is composed mainly of β sheet, which is less successful for molecular replacement studies, than α helical structures.

The success of molecular replacement studies obviously depends on the similarity of the search model and the unknown. Other factors affecting the degree of success are: having an accurate highly refined search model (Huber, 1985) and the completeness of the search model. The completeness of the search model affects the accuracy to which model structure factors can be calculated and the RF value. The type of structure being looked at also influences the way in which the molecular replacement study proceeds: α helical models can be examined at low resolution (e.g. some of the early studies on haemoglobin) since even at 6Å resolution cylinders of density from the α helices give strong Patterson vectors. However, it is often necessary to go to higher resolution, approximately 3Å when looking at predominantly β sheet structures, even though the sequences of these elements might be very different.

```
                        10         20         30         40         50
                     <-βA--> <-----αB------><--βB-----> <-------αC----->
>HSD                 MNDLSGKTVIITGGARGLGAEAARQAVAAGAR--VVLADVLDEEGAATARELGDAAR--Y
>DlADH               M-DLTNKNVIFVAALGGIGLDTSRELVKRNLKNFVILDRVENPTALAELKAINPKVNITF
                     * **..*.**....  *.* ...*. *  . .  *.*. * . .. *. ....  ..  .
                       60         70         80         90        100        110
                     <-βC->     <---αD---->    <----βD---->      <-------αE------
>HSD                 QHLDVTIE-EDWQRVVAYAREEFGSVDGLVNNAGISTGMFLETESVERFRKVVDINLTGV
>DlADH               HTYDVTVPVAESKKLLKKIFDQLKTVDILINGAGI-----LDDHQIER---TIAINFTGL
                     .  ***.  .. ....    ... .** *.*.***     *... .**   ...**.**.
                       120          130        140        150          160
                     ---------------> <----βE-->          <-----------αF---------
>HSD                 FIGMKTVIP---AMKDAGGGSIVNISSAAGLMGLALTSSYGASKWGV----RGLSKLAAV
>DlADH               VNTTTAILDFWDKRKGGPGGIIANICSVTGFNAIHQVPVYSASKAAVVSFTNSLAKLAPI
                       . ....     .*.. ** *.**.*..*. ..  .. *.*** .*    ..*.***.
                    170        180        190        200        210        220
                     ---> <----βF-->     <------αG----->
>HSD                 ELGTDRIRVNSVHPGMTYTPMTAETGIR-QGEGNYPNTPMGRVGNEPGEIAGAVVKLL-S
>DlADH               T----GVTAYSINPGITRTPLVHTFNSWLDVEPRVAELLLSHPTQTSEQCGQNFVKAIEA
                     .     . . *..**.* **.. . . . . * . ..  ... ..... . . ** . .
                      230        240         250
                            <-βG>
>HSD                 DTSSYVTGAELAVDGG--WTTGPTVKYVMGQ
>DlADH               NKNGAIWKLDLGTLEAIEWTK----HWDSHI
                     .... .   .*..  .. **.    ..

Key:   .  = conserved substitution
       * = identity
```

Figure 4–7: Sequence alignment between HSD and DlADH. Sequence aligned using CLUSTAL with default settings. The elements of secondary structure of the HSD are marked <—>

Accuracy of search model

The HSD structure available was determined to 2.6 Å resolution. The chain trace was carried out in one subunit and the other three subunits were generated from the 222 symmetry. The R-factor after refinement with noncrystallographic symmetry restraint, including 7424 protein atoms (255 residues and 1856 atoms/subunit) was 0.231 for 26,753 ($F > 2\sigma_F$) reflections between 6.0 to 2.6 Å, data are 85% complete. The r.m.s. deviations from ideality of some of the geometrical parameters are; 0.025 Å for bond distance; 0.061 Å for angle distance and 3.3° for bond angles. A Ramachandran plot shows six non glycine residues per subunit in the disallowed region of the plot (see Figure 4–8). These residues lie in the loop regions (between αC and βC, βC and αD, βF and αG) and in the carboxyl terminal region which have weak densities and very high temperature factors. Further refinement of the structure is in progress (D. Ghosh, personal communication).

4.2.3 Rotation function: MERLOT

The first HSD model used for molecular replacement was the poly-alanine chain of the Q-axis dimer (see Figure 4–6), with dimensions 65 Å x 62 Å x 28 Å (28 Å in the direction of the Q-axis). The Q-axis dimer is the most likely dimer due to the close contacts at this interface. The size of the unit cell, a=70.6 Å, b=55.8 Å, c=65.8 Å and β=107° constrain the Q-axis dimer to lie with its molecular diad almost parallel to the 2-fold crystallographic axis. The molecular diad for the DlADH crystals, lies at 22.5° to the crystallographic 2-fold axis.

The rotation function calculation was carried out using different resolution ranges, but no solution was consistent for different resolution ranges and the solutions did not orientate the dimer close to the crystallographic 2-fold. The search model was modified to remove loops 202-234 and 237-255. These loops have low density and high B-factors in the HSD structure, also there is low identity with DADH in these regions (see Figure 4–7). Some of the guidelines adhered to in this study are outlined below:

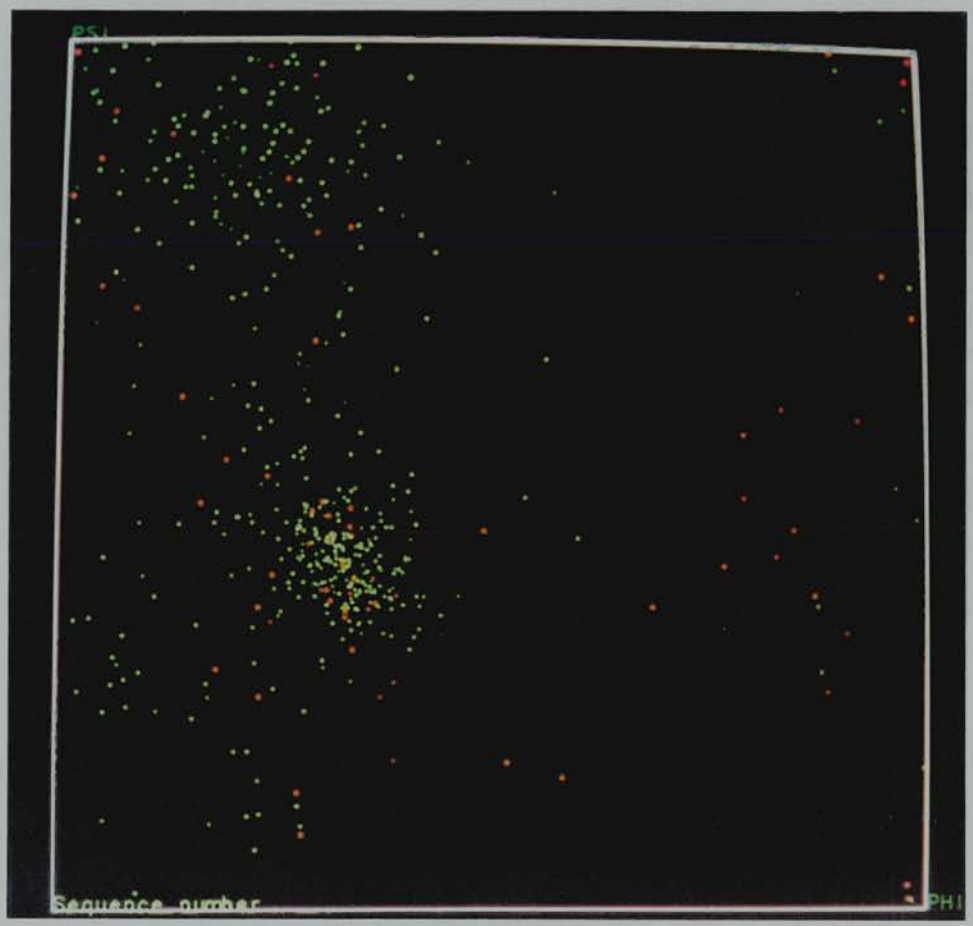


Figure 4–8: $\phi\psi$ plot for the polyalanine chain of HSD.

- The use of high resolution data depends on the similarity of the search molecule to the unknown; generally data of better than 4 Å resolution were not included in calculations.

- Low resolution data contain information about solvent regions and these are different between different crystal forms; they should therefore be excluded from cross rotation calculations.

- The integration radius was chosen to give about 90% integration of Patterson space (Blow, 1985). The maximum radius is limited by the smallest dimension of the unknown molecule.

- Best results have been found with sharpened Patterson functions (Dodson, 1985) where there is an even distribution of F^2 against resolution (this can be done artificially by using data in narrow resolution bands).

- An unrealistic distribution for F_{calc} of the model can be obtained if there are not many atoms in the model. This affects the 3-4 Å resolution data and this could be a problem with using a poly-alanine model.

Initially, CROSUM the Crowther fast rotation function in MERLOT, was used for the cross rotation studies. An integration radius of 23 Å was used with the Q-axis dimer as a search model and an integration radius of 15Å was used when the monomer was used as the search model.

Rotation function solutions were checked by using different data sets. The model without loops gives consistent rotation function solutions for different resolution ranges (see Table 4–1) but not for different data sets. For example, rotation studies carried out on the NATX data (see Table 4–2) do not correlate with solutions found using the CMN data set (see Table 4–1).

The most promising rotation function solution was obtained for the CMN data, the rotation function solution was consistent over all resolution ranges and the solution orientated the dimer diad close to the diad axis found by the self rotation studies. Since the solution to the self rotation solution is known, the

Resolution (Å)	Radius (Å)	Euler Angles α	β	γ	RF
8-4	23	17.5,	102.0,	260.0	4.76
		40.0,	138.0,	295.0	4.67
		102.5,	8.0,	175.0	4.62
10-5	23	17.5,	102.0,	250.0	3.62
		160.0,	84.0,	75.0	3.51
		17.5,	72.0,	285.0	3.50
12-6	23	162.5,	102.0,	155.0	3.83
		157.5,	80.0,	20.0	3.68
		12.5,	104.0,	250.0	3.66

Resolution (Å)	Radius (Å)	Euler Angles α	β	γ	RF	Peak Position
10-5	15	15.0,	99.0,	250.0	3.21	1
10-5	23	17.5,	102.0,	250.0	3.62	1
10-5	26	17.5,	103.0,	245.0	3.45	2
10-5	29	17.5,	103.0,	245.0	3.35	6

Table 4–1: MERLOT Crowther rotation using poly-alanine chain of dimer without loops 202-234 and 237-255 as search model. The cross rotation function was calculated using the CMN data. Results are shown i) as a function of resolution ii) as a function of Patterson integration radius.

Resolution (Å)	Radius (Å)	Euler Angles α	β	γ	RF
8-4	23	22.5,	153.0,	10.0	5.55
		37.5,	26.0,	195.0	5.49
		27.5,	42.0,	20.0	5.26
10-5	23	12.5,	72.0,	20.0	4.94
		2.5,	108.0,	200.0	4.53
		177.5,	40.0,	210.0	4.38
12-6	23	12.5,	72.0,	20.0	4.88
		17.5,	110.0,	155.0	4.76
		5.0,	108.0,	200.0	4.67

Table 4–2: The first three cross rotation solutions for various resolution ranges using the NATX data. Solutions are not consistent for different resolution ranges.

direction cosines for the noncrystallographic diad can be calculated (see POLARRFN program documentation),

$$l = sin\omega * cos\phi \tag{4.25}$$

$$m = sin\omega * sin\phi \tag{4.26}$$

$$n = cos\omega \tag{4.27}$$

where ω and ϕ are as defined for POLARRFN (see Figure 4–1). The POLARRFN solution is ω=22.3° and ϕ=76.5° (κ=180°). These give direction cosines for the diad as, l = 0.0886, m = 0.3690 and n = 0.9252. The self rotation gives the position of the diad but the dimer can be orientated on one of two ways with respect to the noncrystallographic diad: the dimer can lie with its diad parallel or antiparallel to the noncrystallographic diad axis. In an antiparallel position, rotating from 010, the direction cosines are, l= -0.237, m=-0.949, n=-0.208. If the Q-axis diad is initially aligned with the y-axis in the input file then the Euler angles given as a solution to the cross rotation function can be applied to a dummy coordinate file containing the vector 010, the resulting vector will have the direction cosines of the molecular diad. The cross rotation in MERLOT which brought the dimer diad closest to the noncrystallographic diad was found for the CMN data and was α=17.5°, β=102.0° and γ=260.0°.

The accuracy of the cross rotation solution found by CROSUM is improved by using a fine grid search of 0.5° steps and then using a Lattman rotation search (LATSUM in MERLOT). The peak refined to α=18.0°, β=104.0° and γ=256.0° (which orientated the dimer diad so that it has direction cosines l=-0.149, m=-0.303 and n=-0.942). The translation function calculation was carried out for this solution using TRNSUM within MERLOT. TRNSUM uses the Crowther and Blow (1967) formulation and with amplitude and phase component of the model structure factors replaced by just the phase component (Lang, 1985). TRNSUM is a vector search method and problems arise when the molecular symmetry elements lie parallel or nearly parallel to the crystallographic axis. The noncrystallographic diad in the DADH crystals is 22.5° to the crystallographic axis, but there is a possibility that the helices which form part

of the Rossmann fold are nearly parallel with the crystallographic axis, this may cause problems for a vector search translation function. Note that a problem with the MERLOT package is its use of non-standard orthogonalization conventions. The output PDB file from MERLOT had **b*** along Z and **c** along X. Before input into X-PLOR this output had to be rotated so that **b*** was along Y and **axb** was along Z.

The translation function search in X-PLOR was carried out after the rotations solution had undergone 50 steps of rigid body minimization, using the Patterson correlation as the target function. For 4964 reflections between 4.5-9.0 Å resolution (NATX data) the correlation factor after refinement was 0.025. The translation calculation proceeded using 1916 reflections between 5-10 Å resolution. The maximum peak, $\frac{TF}{\sigma_{TF}} = 5.24$ and packing 0.3004 corresponded to a molecular position with x = 0.21 and z = 0.25. The TF value is low compared to examples in the X-PLOR manual e.g. TF = 0.2, but this solution matched with solutions found using TRNSUM, using data between 4-10 Å. Solutions were not consistent for different resolutions nor were they consistent for different data sets. Rigid body refinement gave an R-factor for this solution of 57% for 9080 reflections. The packing of this solution was looked at using FRODO, close contacts between symmetry related molecules are between loop regions of the protein. The solution was further checked by using phases calculated from this molecular replacement solution to calculate difference Fourier maps. Peaks in the difference Fourier synthesis did not correspond to peaks found in the difference Patterson maps. The solution was discarded.

4.2.4 Monomer as search model

Some preliminary studies with the monomer of HSD, with loops 202-234 and 237-255 deleted, were also carried out in MERLOT and in X-PLOR. They did not give consistent peaks for different resolution ranges. In MERLOT studies were carried out for data between 4-8 Å, 5-10 Å and 6-12 Å and with a Patterson radius of 15 Å. The best solution for the 5-10 Å data, moved the monomer close

to the noncrystallographic axis: α=17.5, β=110.0 and γ=305.0. The X-PLOR rotation function carried out for the monomer used the same resolution ranges of data, but with a radius of 17 Å. The solution placed the monomer in a similar orientation to the MERLOT solution but in an anti-parallel fashion with respect to the Q-axis interface.

4.2.5 X-PLOR rotation function

Since all translation functions depend on the accuracy of the rotation solution it was decided to undertake further studies of the rotation function using X-PLOR using the Q-axis dimer as a search model. The X-PLOR rotation function uses a real space Patterson search method. This has proved to be beneficial in a number of cases (Huber *et al.*, 1985) and also incorporates a Patterson correlation refinement. This was used to modify the search model by altering the relative orientation of the monomers and also the relative positions of each element of secondary structure. The refinement of each atomic position in the search model was computationally too expensive to be practical. PC refinement, where each element of secondary structure was moved independently, was tried but this caused helices lying at the dimer interface to penetrate each other because there is no empirical energy term used when rigid body type refinement is used. PC refinement was used to refine the relative orientation of each monomer. With the implementation of X-PLOR on the Digital ALPHA T1.0 VMS SYSTEM (EUCS), the refinement of atomic positions of the search model using PC refinement is now possible.

The cross rotation function in X-PLOR was carried out with all data sets. The best solution did not position the molecular diad near the noncrystallographic diad.

4.2.6 Direct search rotation function

Since the direction of the noncrystallographic diad was known it was possible to orientate the search dimer so that it's molecular diad lies parallel to this axis. The rotation function can then be constrained so that the rotation search is a one dimensional search, about that axis. The dimer was explicitly rotated and then the PC target function was computed for each orientation. The dimer was manipulated using the pseudo-Eulerian angle coordinate system (Lattman, 1985). Where θ_1 and θ_2 define the diad and θ_3 defines the rotation about that diad. A coarse search was carried out with θ_3 moving in 10° steps, this was followed by a finer search about the peak position, using 2° steps. A minimum in the effective Patterson energy term indicates a maximum overlap of the search model Patterson on the unknowns Patterson, for the DADH crystal this lies at, $\theta_1 = 23.0°$, $\theta_2 = 144.0°$ and $\theta_3 = 330.0°$ and has a maximum correlation of 0.0589 (c.f. MERLOT solution correlation of 0.025). The accuracy of the direct search is limited by the manual positioning of the search model. This initial positioning of the model, may introduce inaccuracies into the position of the dimer, however, the rotation solution obtained was refined using PC refinement (after 50 steps of rigid body refinement) which moved the monomers with respect to each other (see Figure 4–9). The direct search carried out using the CMN data sets and both NATX data sets gave the same result although the PC refined solutions were slightly different (see Figure 4–9). The PC refinement was carried out at a resolution 4.5-9.0 Å and reflections with $I > 2\sigma_I$.

4.2.7 Translation function

The translation function calculation was carried out using this PC refined dimer, in all but one case (see Table 4–3) and with several data sets and resolution ranges. Reflections were selected using a 2σ cutoff. Twenty five steps of rigid body refinement were carried out using data between 4.5-9.0 Å resolution, the correlation factor after this refinement is reported. The translation function calculation was carried out using data between 5-10 Å and on a 0.25 Å grid. The

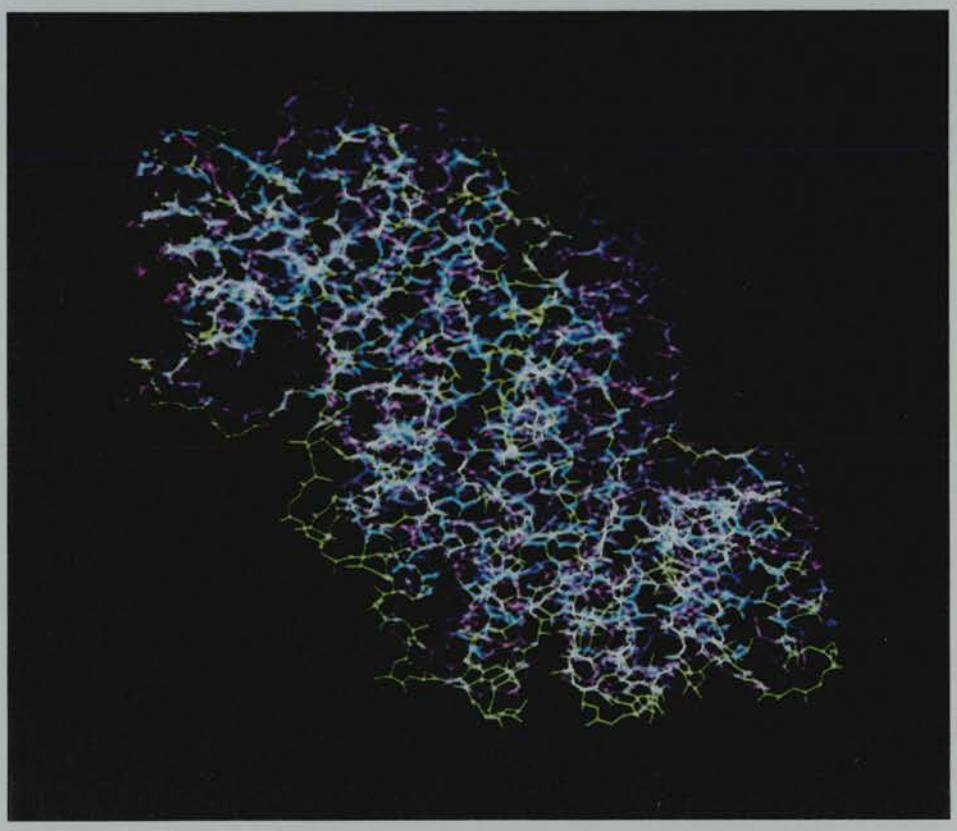


Figure 4–9: Plate showing the rotation solutions after PC refinement with CMN data set (blue), GNAT(XSCALE) data (green) and the GNAT(ROTAVATA/AGROVATA) data (purple).

Data set	Rigid body		Translation function			
	No.ref.	Corr.	No.ref.	$\frac{TF}{\sigma_{TF}}$	Packing	Peak
CMN	2592	0.089	3501	6.94	0.253	0.20, 0.0, 0.14
				6.44	0.253	0.08, 0.0, 0.35
NATX	5141	0.092	1907	7.15	0.251	0.22, 0.0, 0.14
				6.5	0.251	0.30, 0.0, 0.25
NAT	4964	0.077	1916	2.08	0.253	0.49, 0.0, 0.28
				2.04	0.251	0.28, 0.0, 0.26
NAT *	4964	0.064	1916	2.12	0.252	0.14, 0.0, 0.39
				2.04	0.252	0.22, 0.0, 0.13
				2.0	0.248	0.30, 0.0, 0.24

Table 4–3: Table summarizing the translation function results from different data sets. * This model was not subjected to PC refinement before the rigid body refinement and translation function were carried out.

translation function results are summarized in table 4–3. The most consistent and strongest peak in the translation function corresponded to a position x = 0.201 and z = 0.140 along. The translation functions carried out with different resolution ranges of data gave clusters of peaks with the top peak much less than 1σ greater than the second peak.

4.2.8 Direct search solutions

Attempts to refine the direct search rotation and translation solution using rigid body refinement, resolution 3-8 Å failed to reduce the R-factor below 56% (initial starting R-factor $\approx$ 60%). The phases from this MR solution were used to calculate a difference Fourier map using a derivative data set. Heavy atom positions calculated from the difference Fourier map agreed with the heavy atom peaks found on the difference Patterson maps, suggesting that the MR solution was correct. However, the MR map calculated, $3F_o$-$2F_c$ map with data between 20 - 4Å was uninterpretable.

The poor quality of the MR map suggested that the positioning of the search model had not been determined accurately enough (or that the search model was significantly different to the unknown). It was decided to use simulated annealing (SA) in the refinement stage of the molecular replacement solution. The radius of convergence for SA is larger than that of conventional rigid body

refinement, and thus may allow for large positional changes in the molecular replacement solution. SA allows the model to move freely under some X-ray and some geometrical constraints, this can result in significant modifications to the search model and hence the calculated phases.

A further direct search was carried out with the Q-axis dimer model positioned so that the model diad was anti-parallel to the noncrystallographic diad, the rotation solution had $\theta_1 = 21.0°$, $\theta_2 = 153.0°$, $\theta_3 = 330.0°$. This solution gave an R-factor after rigid body refinement of 56.1%. Simulated annealing refined the model to an R-factor = 37.4% with a total energy = 57852.0 kcal/mol. This can be compared to the model aligned with its molecular diad, parallel to the noncrystallographic diad, which had a total energy of 25065.1 kcal/mol and an R-factor of 39.8%; and a randomly orientated search model which refined to have a total energy 26171.2 kcal/mol and an R-factor of 39.3%. A summary of the simulated annealing results is shown in table 4–4.

The packing of the direct search solution (with the Q-axis dimer parallel to the molecular diad, and with a translation of x = 0.20 and z = 0.14) was examined in FRODO (see Figure 4–10). No close contacts between symmetry related molecules were observed. The molecules pack so that there are large solvent channels running perpendicular to z. It is likely that the crystal contacts across these channels are through loops 202-234 and 237-255, the loops that were excluded from the MR study.

4.2.9 Reassessment of MERLOT rotation and translation solution

In the light of the direct search rotation function (see section 4.2.6) the MERLOT rotation solution was reassessed. The MERLOT rotation solution had been found to bring the Q-axis diad into a position that would account for the self rotation function solution. However, the MERLOT rotation function solution is slightly different to the rotation solution found using the direct rotation search (see Figure 4–11). Translation function studies using the MERLOT rotation

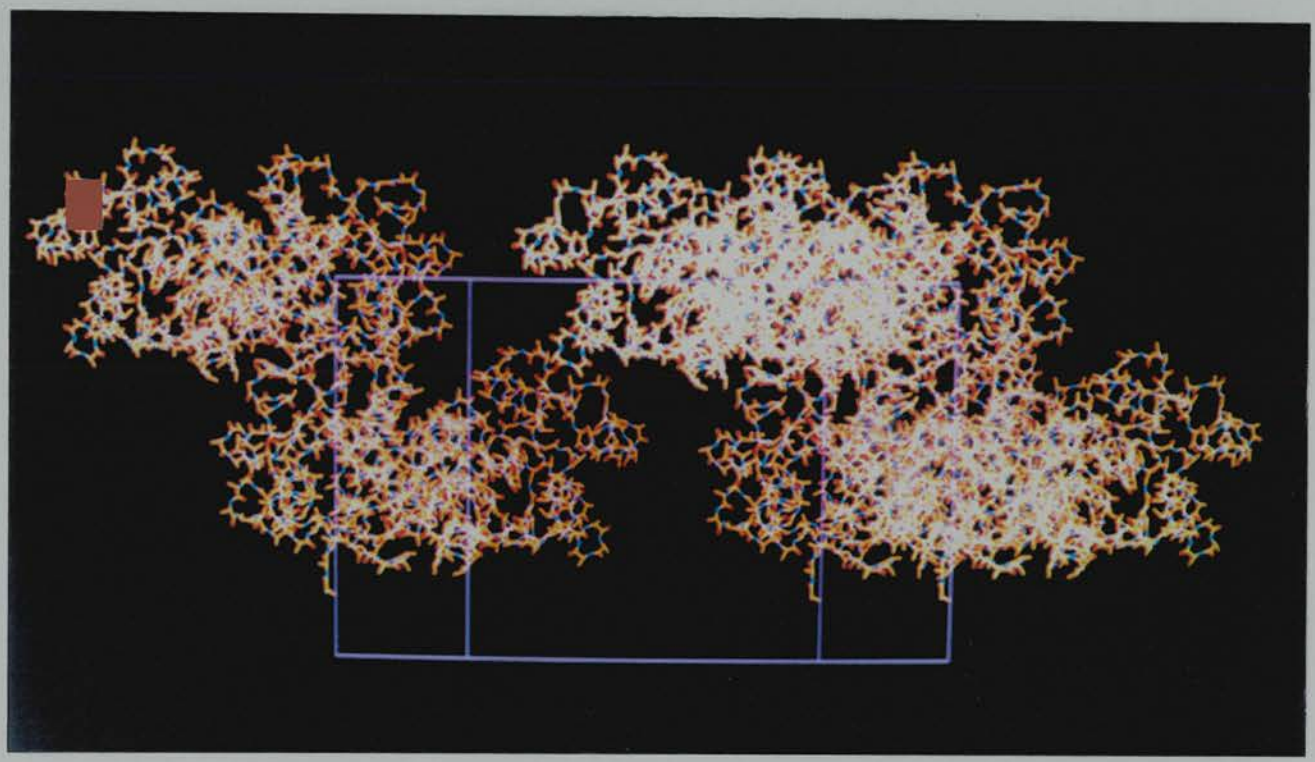


Figure 4–10: Packing diagram DlADH has determined by the molecular replacement solution. The packing of the polyalanine model of the Q-axis dimer is shown, with loops 202-234 and 237-255 removed.

solution had given a number of solutions but only the solution from TRNSUM (found using the CMN data, in the 5-10 Å resolution range) gave a translation solution equivalent to that found for the direct search rotation solution.

4.2.10 Difference Fouriers

A good test for the accuracy of the phases determined by molecular replacement is to use the phases to calculate difference Fourier maps using any heavy atom data. The difference data collected on the 2-chloromercuri-4-nitrophenol soaked crystals (GCMN) and the mercury chloride (GHG) data were used to calculate difference Fourier maps. The phases calculated from the direct search molecular replacement solution for data between 20 - 5 Å were used to compute this difference Fourier map. Although solution peaks corresponding to peaks in the difference Patterson maps were weak they produced cross vectors which were consistent with peaks in the difference Patterson. A molecular replacement solution where the search model was rotated randomly and then moved to the translation solution position (x = 0.20, y = 0.0, z = 0.14) was used to generate phases for the data between 20 - 5Å resolution. Difference Fourier maps computed with these phases did not locate the heavy atom positions. However, it has been reported that an incorrect translation function solution was used to successfully locate heavy atom positions using difference Fourier techniques, in the structure determination of T-state phosphofructokinase (Evans *et al.*, 1985), also that in cases where five of the six positional parameters are correct the phases calculated from this solution will be good enough to locate the heavy atom positions using difference Fourier synthesis (Dodson, 1992).

4.2.11 Refinement of solutions

Simulated annealing refinement was undertaken in an attempt to improve the MR map and obtain further statistics on the MR solutions. The direct search solution, the MERLOT solution and some control solutions were refined using simulated annealing (see Table 4–4) and the results were compared.

Solution	data set	R-factor (rigid body)	E(tot) (after SA)	E(xref)	R-factor
Direct search	NATX data	57.0	25065.12	23522.39	39.8
	CMN data	55.0	18980.32	18163.42	40.7
	NAT data		36607.31	32999.49	38.8
Anti-parallel	NATX data	56.1	57851.98	49219.78	37.4
Merlot solution	CMN data	57.3	28282.10	25446.48	37.8
Random translation					
(0.48, 0.0, 0.28)	NAT data		40758.70	36415.34	38.6
Random orientation	NATX data	56.3	26171.18	24005.91	39.3

Table 4–4: Table shows the R-factor after rigid body refinement and then after annealing. Also the total energy and the X-ray energy after annealing. The direct search solution is refined using GNAT, CMN and NAT data sets. A different translation solution is refined with the same orientation. The same translation solution is refined but using different orientations of the molecule.

It seems that the R-factor is not a good indicator of the correctness of the model. The energy terms are high for the incorrect translation solution but an incorrect orientation does not give a significant change in the energy. It is interesting to note the variation in the energies calculated for the correct solution using the different data sets.

SA improved the MR maps, that is they showed more continuity and less noisy density, although they were still too poor to chain trace. The phases calculated from the annealed MR solution were better for interpreting the difference Fourier maps.

4.2.12 Electron density maps

The electron density maps were calculated using the direct search and MERLOT molecular replacement solutions. Maps were calculated (X-PLOR) with data between 20 - 4 Å and with coefficients,

$$3F_o - 2F_c \tag{4.28}$$

Where $\mathbf{F}_o$ is the observed structure factor with calculated phases and the $\mathbf{F}_c$ is the complex structure factor for the model. This removes some bias that may be introduced by using phases from a model. The electron density map

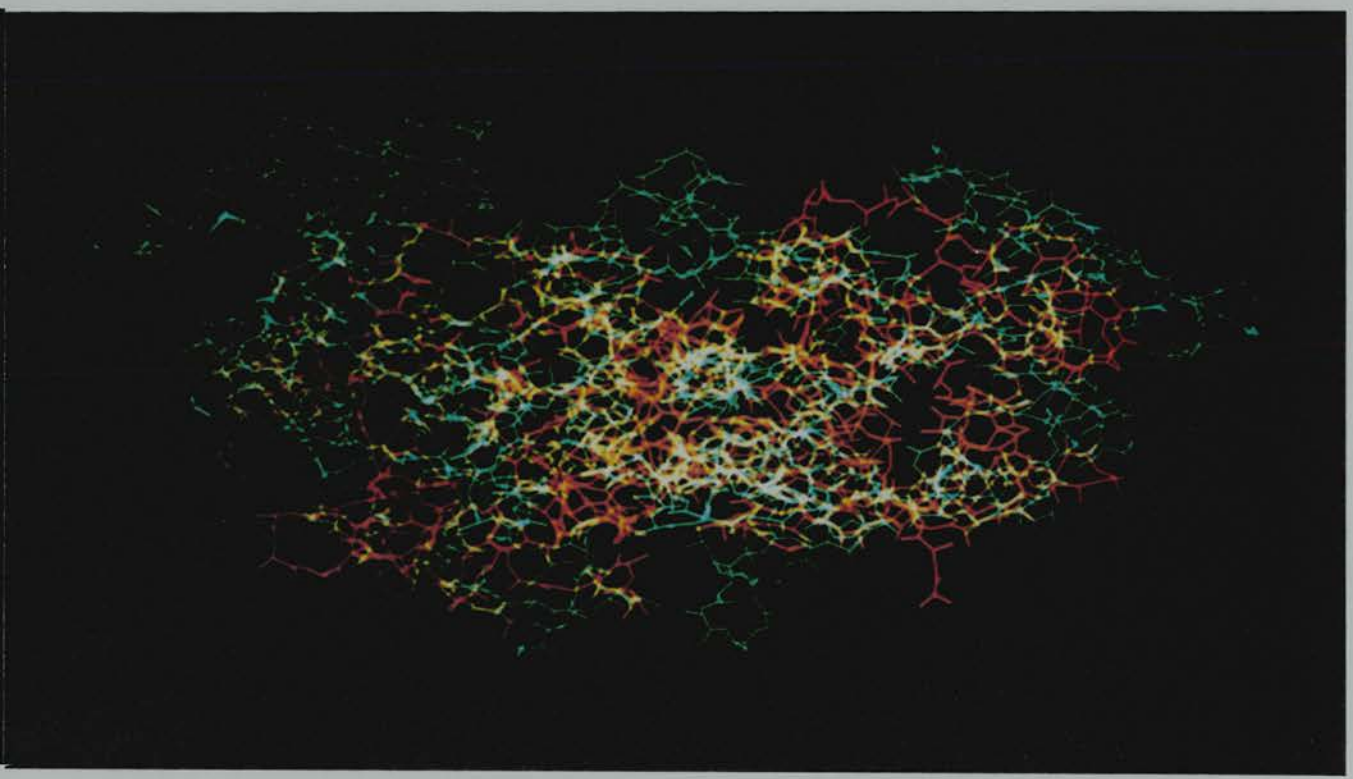


Figure 4–11: Plate showing the molecular replacement solutions from MERLOT (red) and from the direct search method within X-PLOR (green).

corresponding to the six parallel β strands that form the Rossmann fold are shown in figures 4–13 and 4–12. A correct molecular replacement solution should produce a map which reveals features that are not apparent in the search model used to calculate the phases e.g. a correct map for DlADH should show side chain density where there is none in the poly-alanine search model and changes in loop regions where the unknown and the search model are known to differ. There seems to be a lack of detail in the side chain densities. Although both the direct search MR solution and the MERLOT MR solution show good continuity the MERLOT MR map is more noisy.

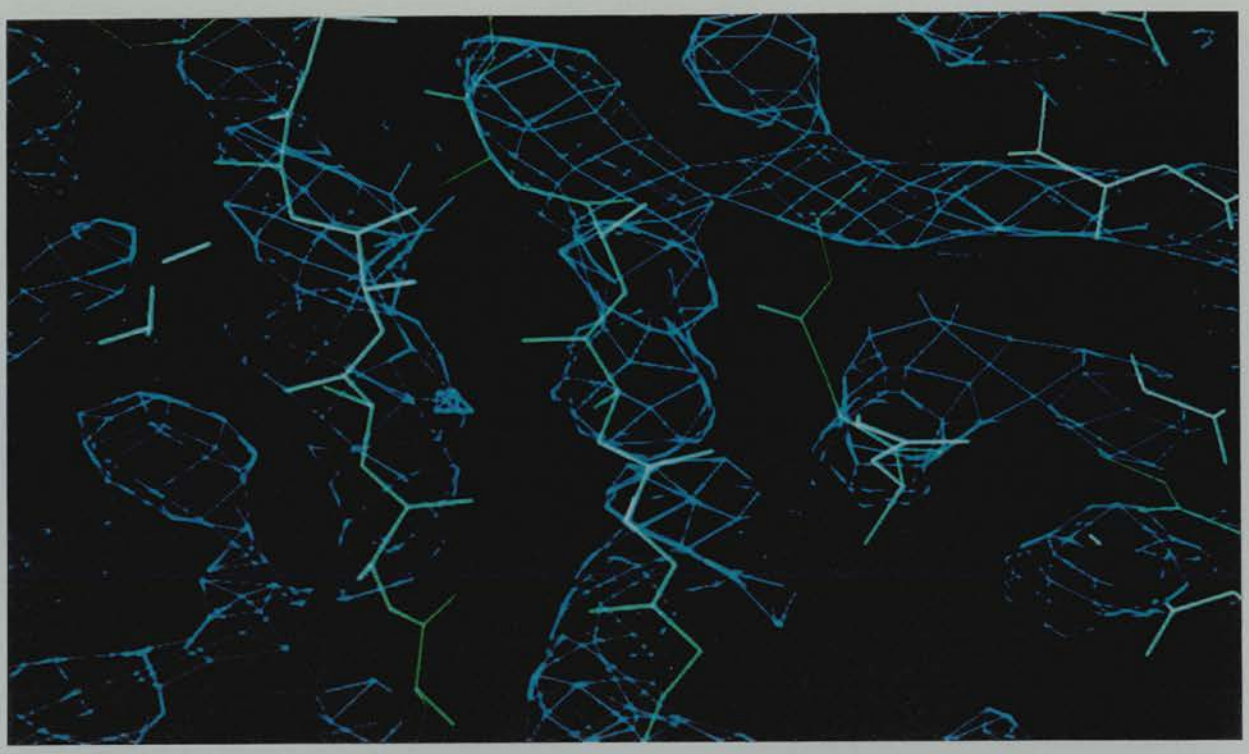
4.3 Discussion

Self rotation studies were carried out on data indexed in both the small and large cells. In the small cell there is one dimer per asymmetric unit, the two fold axis in this dimer has ω=22.3°, ϕ=76.5°, which orientates the molecular dimer at 20° to the **b** axis.

The best cross rotation function solution was found using the direct rotation search in X-PLOR where the dimer axis was fixed so the rotation function was a simple one dimensional search. The rotation function solution found by this method was consistent for all data sets and for all resolution ranges. Translation function studies for this orientation gave a significant solution peak with data between 5-10Å for all data sets. Other resolution ranges did not give significant peaks. The phases calculated from this MR solution were used to calculate heavy atom positions using difference Fourier techniques. The electron density maps calculated from this molecular replacement solution were of poor quality, the 20 - 4 Å map showed regions of continuity and α helices and β sheets are visible. However, in parts the map is noisy and there is a lack of side chain detail which makes map interpretation impossible at this stage. Poor map quality is probably due to:

- The poor phasing power of a poly-alanine model

a)



b)

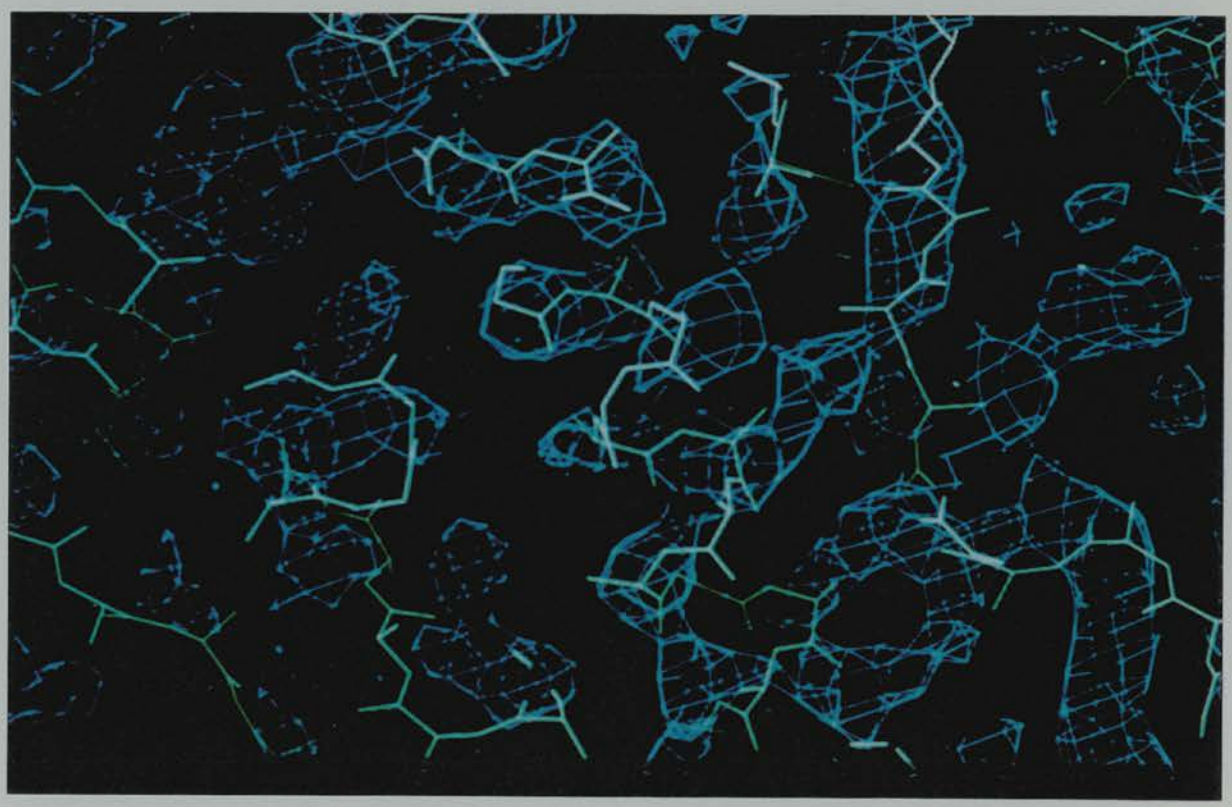


Figure 4–12: Plate showing the electron density map ($3F_o$ - $2F_c$), resolution 20 - 4 Å, from the direct search molecular replacement a) a section of β sheet b) a section of α helix.

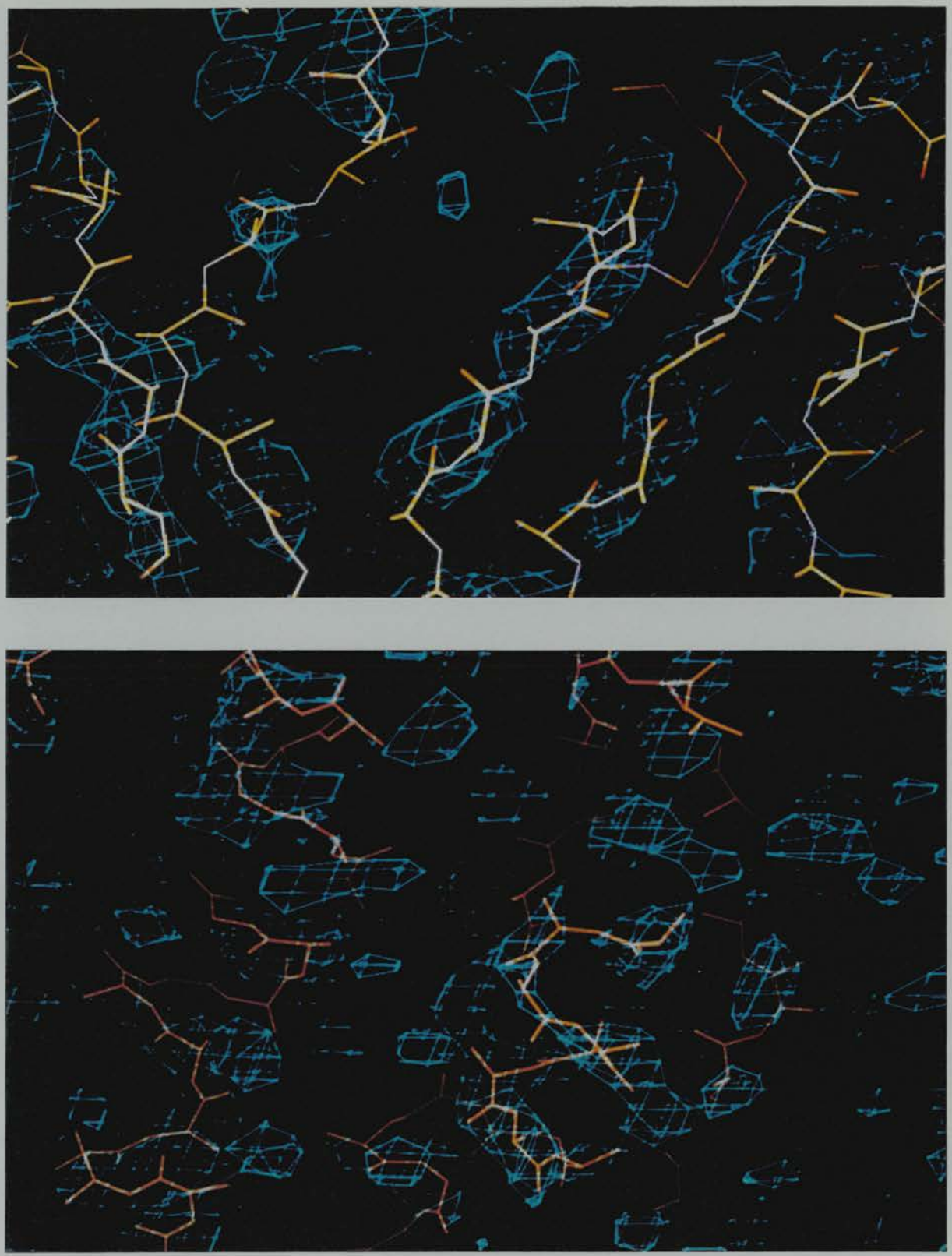


Figure 4–13: Plate showing the electron density map (20 - 4 Å) from the MERLOT molecular replacement a) shows a region of β sheet and b) shows a region of α helix.

- Error in positioning search model for direct search
- Error in positioning poly-alanine model accurately, although for a one dimensional search this error is much less than that encountered in a three dimensional search
- Differences between the search model and the unknown structure

Refinement of the direct search MR solution was used to try and improve the map quality. SA is capable of introducing large changes in the search model coordinates. A disadvantage with this method is that it might cause severe distortions in the HSD structure to try and account for the observed electron density. The map calculated from MR solution after SA, was less noisy than the initial map but it was still not good enough to chain trace.

The rotation function studies within MERLOT seem to be aggravated by inconsistencies between data sets. In contrast, the direct search method, gives a rotation function solution which is reproducible for all resolution ranges and data sets tested. The disadvantage of the direct search method is the inherent inaccuracy of the starting position; a position which has been dependent on identification of the molecular diad axis by eye. Even so, this direct rotation search solution has been used to determine a translation solution which is consistent for different data sets and gives a sensible packing arrangement (see Figure 4–10).

Checking the validity of a promising molecular replacement solution is always a problem. In this case, a problem is that most of the molecular replacement statistics are misleading since a poly-alanine chain is being fitted to electron density from a 'real' protein structure in an attempt to overcome these problems numerous control experiments were run in tandem with the main study. It seems that the R-factor is a poor indicator of solution quality. The correlation function calculated at various stages seems to indicate a better result as does the TF/σ_{TF} value. At the refinement stage, a low X-ray energy term after SA may indicate a correct solution. In this study, most emphasis has been placed on attaining

reproducible solutions using different programs, resolution ranges of data and data sets. Further assessment of correctness has involved using phases calculated from this molecular replacement solution to locate heavy atom positions using difference Fourier synthesis. (Refinement and phasing of these heavy atom positions is discussed in chapter 5).

Further molecular replacement studies for the DlADH data should not be undertaken without access to a more complete and highly refined search model.

Chapter 5

Isomorphous Replacement

5.1 Introduction

This chapter describes the preparation and analysis of isomorphous heavy atom derivatives for the crystals of ADH from *Drosophila*. Results are included for the data collection, scaling and analysis of the data from these crystals, the location of heavy atoms and the refinement of the heavy atom positions.

The determination of macromolecular structures *ab initio*, depends on the success of determining phases from diffraction data using the method of isomorphous replacement (IR). This technique was first used for the structure determination of haemoglobin (Green *et al.*, 1954; Blow, 1958). A full explanation of the theory of the method and its application can be found in Blundell and Johnson (1976).

5.1.1 The phase problem and isomorphous replacement

If the structure factor amplitudes and their phases are known then the electron density in a unit cell can be calculated. However, observed crystallographic data consist of structure factor amplitudes only. It is therefore necessary to obtain an estimate of the phases for these structure factors so that an initial electron density map can be calculated.

The phases of the structure factors are a function of the positions of the scatterers, that is the electron density distribution, in the unit cell. By replacing disordered water molecules in the native protein crystal with a few heavy atoms, in specific sites, it is possible to change the diffraction pattern. The heavy atoms diffract X-rays more strongly than light atoms e.g. carbon, oxygen and nitrogen. When this heavy atom addition leaves the crystal lattice unchanged, this derivative crystal is said to be perfectly isomorphous to the native crystal. In this case, the derivative structure factor is a vector sum of the heavy atom structure factor and the native protein structure factor.

The change in the observed intensities from a crystal as a result of heavy atom

addition can be calculated by:

$$\Delta I \approx Z(\frac{N_H}{N_P})^{1/2}(\frac{f_H}{f_P}) \tag{5.1}$$

(Crick and Magdoff, 1956) where N_H is the number of heavy atoms in the unit cell, N_P is the number of protein atoms in the unit cell, f_H is the average scattering factor for the heavy atom and f_P is the average scattering factor for the protein atoms. For centric reflections, Z=2 and for acentric reflections $Z = \sqrt{2}$. Larger changes in intensity can be achieved by adding more heavy atoms, but this is likely to result in the derivative crystal being non-isomorphous with the native crystal and it also makes locating the heavy atom positions more difficult. Substitution by a single heavy atom can give significant changes in intensities, it is easy to locate and the derivative crystal is more likely to be isomorphous with the native crystal. If the derivative crystal is isomorphous the differences between the two diffraction patterns arise from the substituted heavy atoms only, however, there are always some differences due to non-isomorphism.

The positions of the heavy atoms can be determined using methods that do not require any phase information. Once the position of the heavy atom is known, they can be used as a phasing model from which other atomic positions can be developed. Using the cosine rule, the relationship between the derivative crystal structure factor, ($F_{PH}e^{i\alpha_{PH}}$), the heavy atom structure factor, ($F_He^{i\alpha_H}$), and the native protein crystal structure factor, ($F_Pe^{i\alpha_P}$) can be described (see Figure 5–1):

$$F_{PH}^2 = F_P^2 + F_H^2 + 2F_PF_Hcos(\alpha_P - \alpha_H) \tag{5.2}$$

This can be rearranged to give the protein phase for each reflection:

$$\alpha_P = \alpha_H + cos^{-1}(\frac{F_{PH}^2 - F_P^2 - F_H^2}{2F_PF_H}) \tag{5.3}$$

where α_P is the phase of the protein structure factor; F_P is its magnitude; α_H is the phase of the heavy atom structure factor; F_H is its magnitude. F_{PH} is the magnitude of the derivative structure factor. Since the cosine is a symmetrical function, there are two possible values for α_P. Addition of information from a number of derivatives will give an unambiguous determination of the phase. This

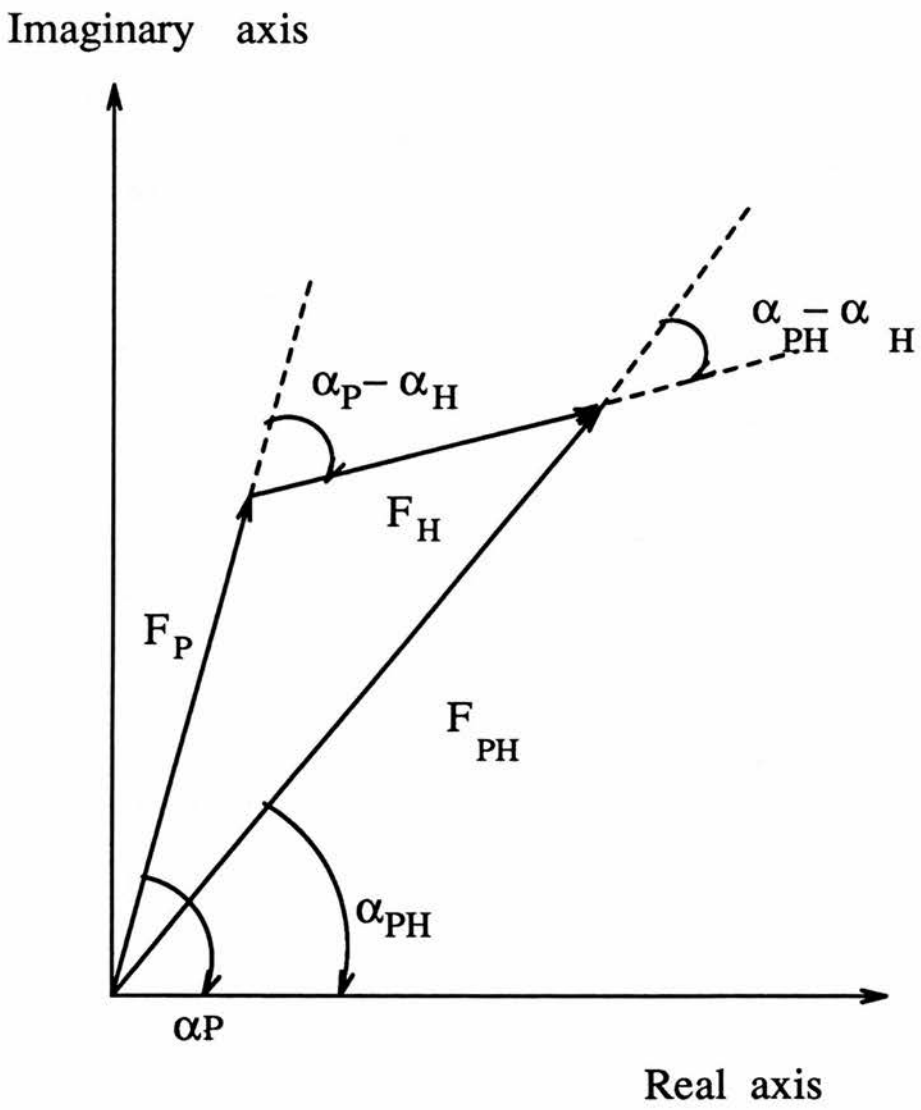


Figure 5–1: Argand diagram showing relationship between $\mathbf{F}_P$, $\mathbf{F}_{PH}$ and $\mathbf{F}_H$.

can be seen clearly in a schematic representation called a Harker Construction (see Figure 5–2). Due to errors in the amplitudes of F_P and KF_{PH} (where K is a scale factor) and errors in the heavy atom model, the phase triangle (compare Figure 5–1 and Figure 5–3) is always distorted which leads to errors in the protein phase. The phase probability distribution for a single derivative is bi-modal (Blow and Crick, 1959). It is usual to combine the phase probabilities from several derivatives and also to include anomalous scattering data whenever possible (Hendrickson and Lattman, 1970).

Anomalous scattering is an effect observed when the frequency of the incident X-rays is comparable to the resonant frequency of the inner electrons of an atom (the absorption edge of the atom). These conditions mean that the incident X-rays are strongly absorbed by the atom. The energy from the absorbed radiation either excites an electron to a higher quantum state or ejects it from the atom. In either case, the scattering factor of that atom becomes complex and it can be written:

$$f = f_i + \Delta f_i' + i\Delta f_i'' \tag{5.4}$$

where f_i' is the real part of the anomalous scattering factor and f_i'' is the imaginary part. The $i\Delta f_i''$ term becomes large if the degree of absorption is large. Scattering by an atom causes a phase change of π relative to the incident radiation. An anomalously scattered wave has an additional phase change which results in it being $\pi/2$ in front of the isomorphously scattered wave. The consequence of this phase change is that Friedel's law (which states that the $|F^2|$ values for centrosymmetrically related reflections are equal) breaks down so that, $F(hkl) \neq F(\overline{hkl})$ (see Figure 5–4). Therefore, the intensity differences due to the anomalous scattering may be used to determine the phase unambiguously from a single derivative crystal. The anomalous effect for light atoms is negligible at the wavelength commonly used for protein crystallography. However, for heavy atoms the anomalous signal becomes measurable and useful (especially when using experimental systems where the wavelength of the incident radiation can be tuned to the absorption edge). Although the magnitude of the anomalous signal is small and the errors in measurement are comparably large, The

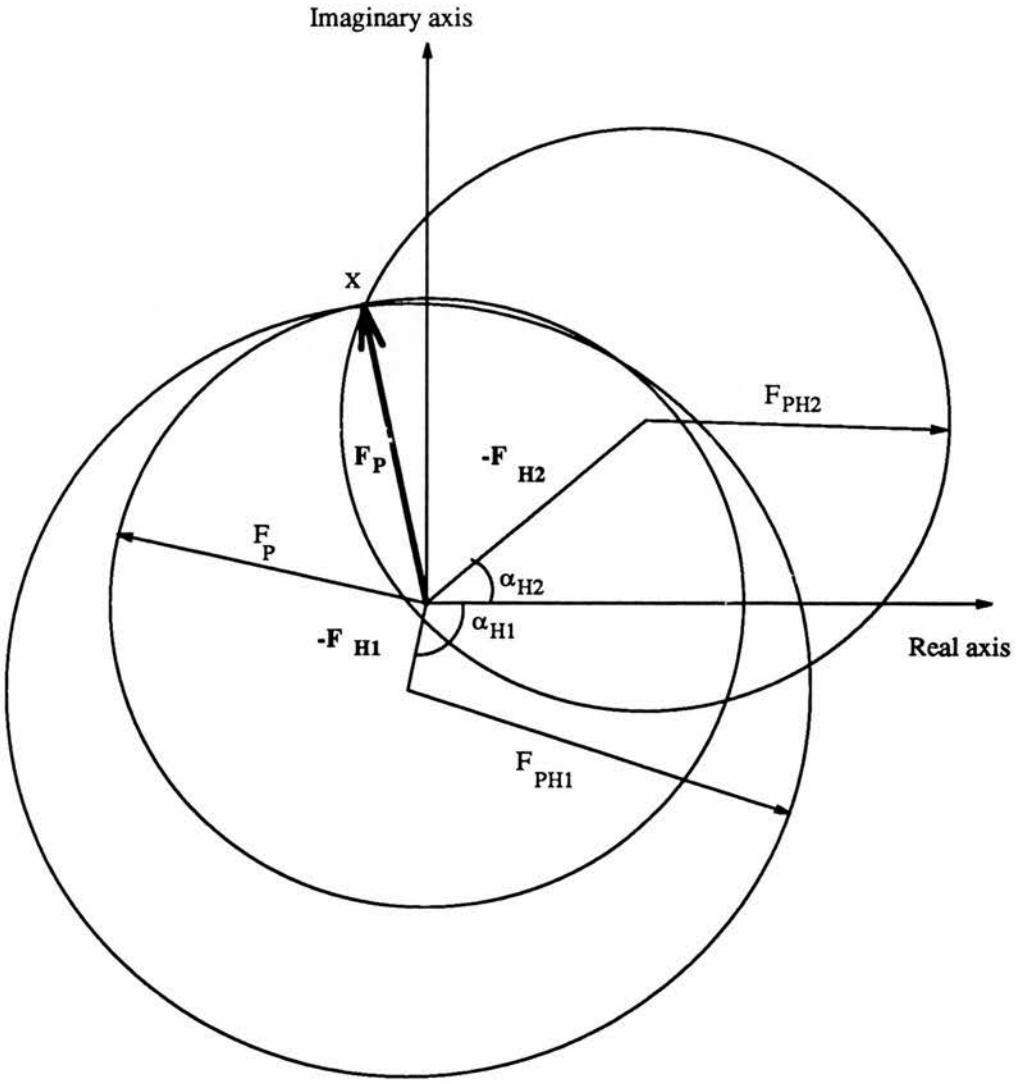


Figure 5–2: The Harker construction representing the phase determination for double isomorphous replacement. The protein phase is unknown but the magnitude of the vector (F_P) is known. The magnitude and direction of the heavy atom vectors $\mathbf{F}_{H1}$ and $\mathbf{F}_{H2}$ are known but only the magnitudes, F_{PH1} and F_{PH2}, of the derivative structure factors are known. It can be seen that using information from both derivatives will give an unambiguous estimate of the protein phase. A single isomorphous replacement experiment would lead to an ambiguity in the protein phase.

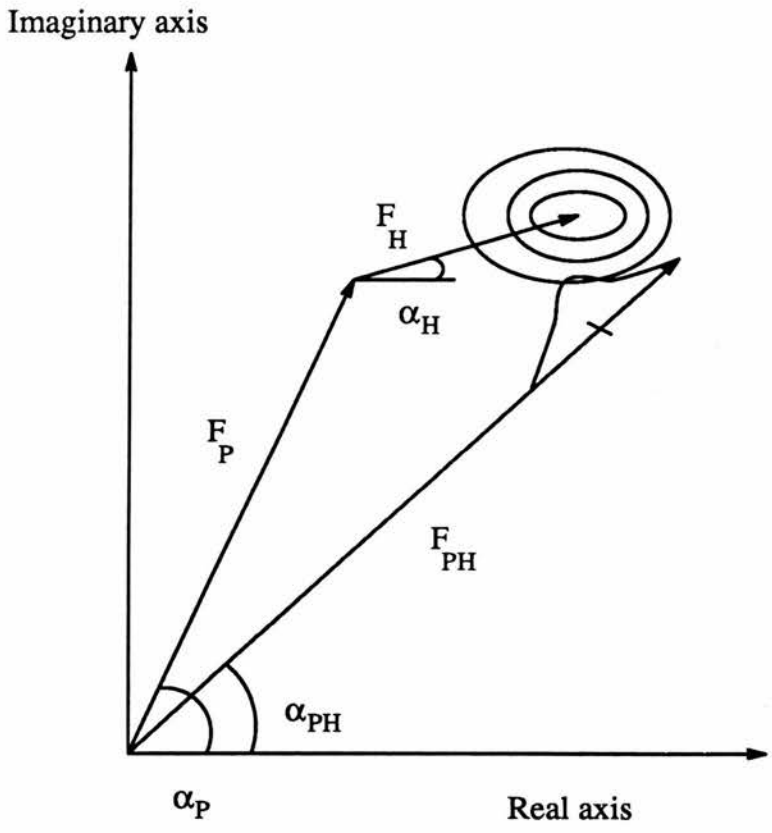


Figure 5–3: Errors in the isomorphous replacement method lead to a distortion and lack of closure of the phase triangle. The error treatment of Blow and Crick (1959) regard all errors as lying in F_{PH}.

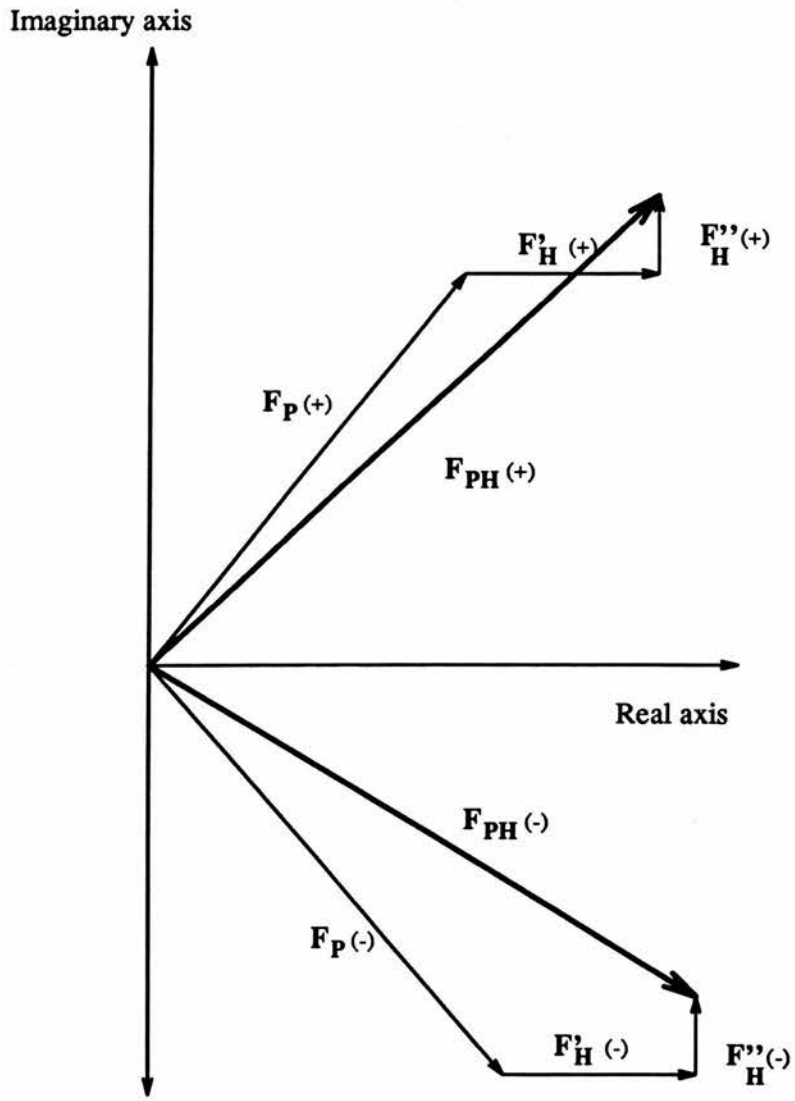


Figure 5–4: Vector diagram showing the effect of anomalous scattering.

advantages of the method are that $F_{PH}(+)$ and $F_{PH}(-)$ are measured from the same crystal so there is no error due to non-isomorphism. Systematic errors in the data collection due to absorption or radiation damage can be reduced by recording $F_{PH}(+)$ and $F_{PH}(-)$ on the same image; this makes the anomalous signal intrinsically more accurate than the isomorphous signal. The anomalous signal comes from the tightly bound inner electrons, therefore it is similar to scattering from a point i.e. the effect becomes proportionally greater with increasing resolution because the normal atomic scattering factor falls with increased resolution. Most often the anomalous signal is used in conjunction with the isomorphous signal where it can be useful in determining the correct handedness of a structure (Bijvoet, 1927; Blundell and Johnson, 1976) and in heavy atom refinement (Dodson, 1976).

Recently, multiwavelength anomalous diffraction experiments have been used to determine several protein structures. Here the anomalous signals, from several wavelengths near the absorption edge, are used to determine the phases (Smith, 1991).

5.1.2 Preparing heavy atom derivatives

There are several comprehensive chapters covering the chemistry of heavy atoms binding to proteins (Blundell and Johnston, 1976; Petsko, 1985). However, the process is still basically empirical. There are several types of heavy atom binding to proteins:

- Metal ion replacement in metalloproteins.
- Heavy atom analogues of specific inhibitors.
- Replacement of amino acids with heavy atom analogues e.g. introduction of selenium methionine into recombinant proteins (Hendrickson *et al.*, 1990).
- The use of site directed mutagenesis to introduce alternative heavy atom binding sites (Nagai *et al.*, 1991)

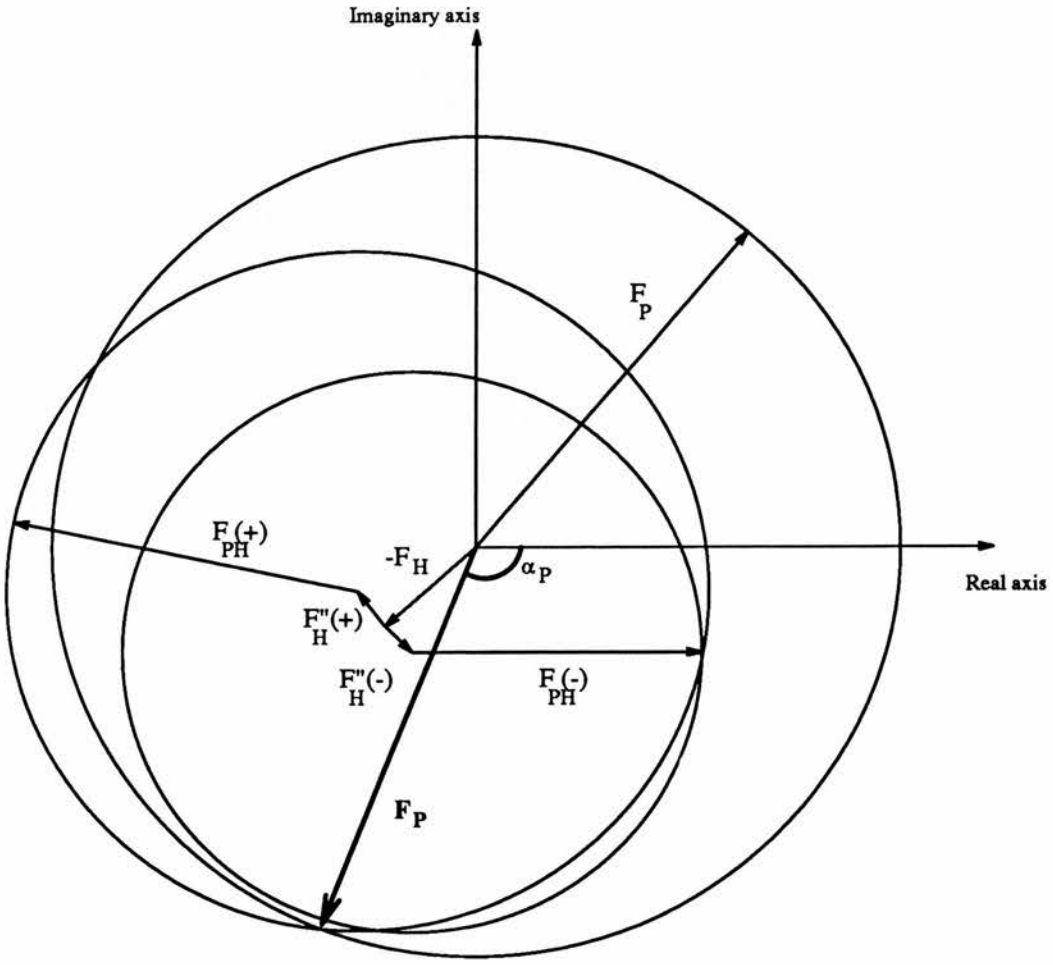


Figure 5–5: Harker construction for anomalous diffraction.

- Direct binding of the heavy atom salts

The methods employed in preparing isomorphous derivatives of DlADH crystals are discussed below:

Heavy Atom Analogues of Specific Inhibitors

The use of heavy atom analogues of specific inhibitors has been used in several structure determinations. However, the addition of a substrate or inhibitor often induces a conformational change in the protein. Also it can limit the degree of useful information about the active site since the binding of an inhibitor can distort the real arrangement of the side chains in the active site.

A heavy atom 'analogue' of NAD^+ was used to determine the structure of horse liver alcohol dehydrogenase (Gunnarsson *et al.*, 1974; Eklund *et al.*, 1974) and since DADH binds NAD^+ it is a possible method of preparing heavy atom derivative crystals. Auricyanide ($Au(CN)_2^-$) and platicyanide ($Pt(CN)_4^{2-}$) anions were cocrystallized with the horse liver alcohol dehydrogenase enzyme. The auricyanide anion was found to bind to two places in the cofactor binding region; the phosphate binding and the adenosine binding sites. The platicyanide ion was found to bind only at the phosphate site.

Direct binding of heavy atoms

The most commonly used method for substituting heavy atoms is direct binding of the heavy atoms to the protein. Heavy atoms can have specific, covalent interactions with amino acid side chains or less specific, electrostatic interactions. Metal ligands can be classified as two types: hard ligands, which tend to interact electrostatically, and soft ligands which tend to form covalent bonds. These ligand types are termed class *a* and class *b* ions respectively. Problems in predicting the behavior of these ligands are due to the binding of the metal complexes instead of binding of the metal ion. For example, soft ligands like Pt, Au and Hg form covalent anionic complexes that interact electrostatically (e.g.

$Au(CN)_2^-$ and $Pt(CN)_4^{2-}$). Electrostatic interactions involve positively charged side chains and charged pockets (e.g. substrate or co-factor binding sites) of the protein which react with heavy atom anions. This type of binding can be affected by the presence of halide ions or phosphate ions in the buffer (see later). The binding of heavy atoms to proteins can be very complicated since protein ligands can form chelating systems which bind heavy atoms non-specifically.

Covalent interactions are the most useful of the heavy atom-protein interactions and the majority of soaking experiments involve covalent binding to the protein of either mercury, silver, platinium or palladium (see later).

Heavy atom derivatives can be prepared by either cocrystallizing protein with heavy atoms or by soaking pre-existing crystals in heavy atom solutions. The disadvantage of preparation by cocrystallization is that the crystals grown may be nonisomorphous with the native crystals, they may even crystallize in another space group. Most heavy atom derivatives are made by soaking pre-existing crystals in heavy atom solutions. Factors that affect heavy atom binding to protein are:

- primary structure of protein
- composition of mother liquor
- pH
- temperature
- concentration of heavy atom solution
- soaking time
- heavy atom complex used

Composition of mother liquor and pH: Ions in the crystal mother liquor react with heavy atoms often forming insoluble complexes. Many heavy atom complexes are sensitive to pH so this must be considered when planning soaking

experiments. Also, ions in the mother liquor may replace ligands in the heavy atom complex. Other additives in the buffer e.g. DTT will alter the reactivity of some heavy atom substitution reactions.

Heavy atom concentration: A high excess of heavy atom reagent is commonly used. The equilibrium of the soaking is dependent upon the relative concentration of the protein to the heavy atom solution. The protein crystal can be regarded as a concentrated solution of protein. When high occupancy sites are not attained the obvious remedy is to increase the concentration of the heavy atom solution, however the risk is that this will increase the number of binding sites and hence lead to nonisomorphism.

Time and temperature: The soaking time required can vary from hours to months. The time required to reach equilibrium depends on:

- the relative size of crystal pores to the reagent
- the time taken for protein to accommodate the heavy atom (i.e. if a slight conformational change of protein is required)
- the reactive species of metal complex (e.g. the lability of the complex)

A short soaking time allows binding to a few fast reaction sites and the chances of maintaining an isomorphous crystal are greater. A long soaking time may give more sites with higher occupancies but the risk of nonisomorphism is greater. However in a few cases, time allows the heavy atom compounds to change and to become more reactive. Platinum compounds react slowly with the buffer so different reactive species will be available in the buffer at different times (Petsko *et al.*, 1978). The protein can also change over a period of time e.g. the deamination of asparagine, or glutamine, or the oxidation of free sulphydryls.

Covalent interactions

Covalent reactions are the most specific and useful of the heavy atom-protein interactions. The covalent interactions used in preparing isomorphous derivatives of DlADH are discussed below:

Mercury

Cysteines, histidines and methionines can all form covalent interactions with class *b* metals. The free sulphydryls of cysteines are often targeted for heavy metal substitution. The most reactive sulphydryls belong to the cysteines, even at pH 6 when the sidechains are almost fully protonated, the reaction with mercury can be rapid. Methionines rarely bind mercury since they have no negative charge and histidines are good ligands at pH 6-7 or greater. Mercury chloride, acetate and nitrate complexes dissociate easily and the mercury atoms can then react with negatively charged or polarizable sulphur ions. However, ammonia and chloride ions in the buffer form complexes with mercury so that it will not react or will only react slowly. Covalent complexes of mercury do not dissociate readily and therefore react slowly or form anions that react electrostatically. The specificity of the mercury complexes depends on size, shape and substituent group, this can be exploited to give good heavy atom derivatives.

Platinum

Platinum is another class *b* metal. $PtCl_4^{2-}$ is one of the most successful compounds in isomorphous replacement experiments. In certain buffers, the platinum compounds gradually change their ligands so their reactivity changes (Petsko *et al.*, 1978); $PtCl_4^{2-}$ can form complexes with the phosphate in the buffer, eg. $[PtCl_3(PO_4)]^{4-}$. This complex is more reactive to protein ligands. $PtCl_4^{-2}$ reacts *via* an S_N2 mechanism, therefore the rate of the reaction depends on both the nature of the leaving group and of the attacking group. This in turn depends on the composition and pH of the buffer. At pH 7 the order of reactivity for the attacking group is

$$\text{imidazole} \trianglerighteq \text{thioether} \trianglerighteq \text{disulphide} \trianglerighteq \text{RSH(RS)} \triangleright \text{RNH}_3^+(\text{RNH}_2)$$

the order of reactivity for the leaving group is less pH dependent

$$H_2O \rhd Cl^- \rhd I^- \rhd NH_3 \rhd CN^- \rhd NO_2^- \rhd CN^-$$

Iridium and osmium Iridium and osmium form stable anions such as $IrCl_6^{3-}$ which can bind through nucleophilic substitution to imidazole or amino groups. Osmium, usually present as OsO_4, is reduced to $OsO_2(OH)_2$ in most crystallization media.

5.1.3 Data collection

In the past, initial screening of derivatives was carried out where possible, using precession photography. In this way, it is possible to obtain an undistorted picture of the reciprocal lattice on a photographic film, which means that native and derivative can be compared optically. A suitable isomorphous derivative crystal has the same cell dimensions (within 1%) and symmetry as the native crystal. The intensity patterns of both crystals should be similar but the intensity changes due to the heavy atom should be evident. Changes due to nonisomorphism are more apparent at high resolution. When crystals are too small and precious to screen by precession photography, as was the case for the DADH crystals, data were collected using an area detector. For derivative data, completeness and accuracy of the data is desirable but not essential. Generally heavy atom derivative crystals are more sensitive to radiation damage, so this must be taken into consideration when planning a data collection strategy.

If the anomalous signal is to be used, it is advisable to align the crystal so that the Bijvoet pairs are collected on the same or adjacent frames; this reduces errors due to absorption and radiation damage. The strength of the anomalous signal can be enhanced, where possible, by tuning the radiation used to the absorption edge of the substituted metal (an advantage of using synchrotron radiation). The anomalous signals should be collected routinely, even at λ=1.54Å when the anomalous signals from most heavy atoms will be weak, since

the information it provides is used in locating heavy atom positions and in refinement of heavy atom parameters (Dodson, 1976).

5.1.4 Scaling and analyzing derivative data

Once a derivative data set has been collected, it is necessary to put the observed intensities on the same scale as those of the native data since the phases are calculated from the relative differences. The CCP4 program SCALEIT, was used to scale the derivative data to the native data.

Initial scaling assumes that:

$$< F_P >=< F_{PH} > \tag{5.5}$$

However, there is a significant contribution from a single heavy atom to the mean scattering intensity which means that the above equation is inaccurate. The data collected on heavy atom soaked crystals were scaled to the native data set using a scale factor and an anisotropic temperature factor. Weak data (where $\frac{I}{\sigma_I} < 3$) were excluded from the scale factor calculation and refinement. An initial scale factor is calculated from the expression:

$$K_{init} = \sqrt{\frac{\Sigma F_P^2}{\Sigma F_{PH}^2}} \tag{5.6}$$

The scale and anisotropic temperature factors are then refined using a modification of the Fox and Holmes method (Fox and Holmes, 1966). The scaled data are used to calculate a difference Patterson, which is relatively insensitive to errors in scaling and therefore an approximate scaling factor can be used. The scale factor is later refined (in MLPHARE) when a model for the heavy atom scatterers is available. Also after scaling, an analysis of the differences between the native and the derivative gives an indication of the quality of the derivative. The quality of the SCALEIT analysis may be determined from several parameters:

R_{iso} plots

R_{iso} is defined as

$$R_{iso} = \frac{<|F_{PH} - F_P|>}{< F_P >} \tag{5.7}$$

A plot of R_{iso} *versus* resolution increases with increased resolution. This is due to random noise and this alone is not a good enough indication of nonisomorphism. However, a sharp rise in R_{iso} with increased resolution might indicate non-isomorphism. The R_{iso} plot has a peak at 7 Å in principle this peak is more pronounced if the changes between the native and derivative are due to a heavy atom substitution rather than to nonisomorphism. Large values for R_{iso} do not necessarily indicate a good derivative nor do small changes indicate a poor one.

Scale factor plots

The scale factor, used to scale derivative to native, when plotted as a function of h, k, l and $\frac{sin^2\theta}{\lambda}$ should be a straight line. Changes of greater than 30% in the scale factor indicate unacceptable nonisomorphism and data beyond this resolution should be discarded.

D_{iso} plot

D_{iso} is defined as:

$$D_{iso} = (|F_{PH} - F_P|) \tag{5.8}$$

A plot D_{iso} as a function of resolution should decrease as a function of increased resolution, in the same way as the scattering curve from a heavy atom. D_{iso} is consistently overestimated due to the errors in measurement of both the native and derivative data sets.

The normal distribution

Although, the normal distribution analysis has been used extensively in small molecule crystallography, it has only recently been applied to macromolecular

crystallography (Howell and Smith, 1992). This procedure compares two data sets to see if they differ systematically. If there is no systematic variation then the observed measurements will all follow a normal distribution of errors. One assumption that the procedure makes is that the standard deviations of the data are estimated accurately. The normal probability-plot based on a plot of $\delta m(real)$ *versus* $\delta m(expected)$, where

$$\delta m(real) = \frac{(|F_1 - KF_2|)}{[\sigma^2(F_1) + K^2\sigma^2(F_2)]^{1/2}} \tag{5.9}$$

and $\delta m(expected)$ can be calculated for any distribution, assuming that the errors have a normal distribution of errors. It has been shown (Howell and Smith, 1992) that for marginal cases, when comparing native and derivative data, the R_{iso} might indicate the presence of a heavy atom, the normal probability plots can distinguish a marginal derivative and a non-derivative. It is proposed that the presence of a heavy atom is indicated by the probability plot having a slope in excess of 15 and an intercept greater than 1.

Difference Patterson maps

The best indication of derivative quality is given by the isomorphous and anomalous difference Patterson maps (see Section 5.1.5). Peaks in the anomalous difference Patterson should coincide with peaks in the isomorphous difference Patterson map, although anomalous difference Pattersons are often noisy and subject to error.

5.1.5 Location of heavy atoms

There are several methods for locating the heavy atom sites in proteins. The methods used in this work are discussed below.

Difference Patterson methods

Phillips (1966) analysed the components that make up the Difference Patterson:

$$\Delta^2_{iso} = (|F_{PH} - F_P|)^2 \tag{5.10}$$

$$\begin{aligned}\Delta^2_{iso} = 4F_P^2 sin^4(\frac{\alpha_P - \alpha_{PH}}{2}) \\ +F_H^2 cos^2(\alpha_{PH} - \alpha_H) \\ -4F_P F_H sin^2(\frac{\alpha_P - \alpha_{PH}}{2}) cos(\alpha_{PH} - \alpha_H)\end{aligned}$$

The first term gives the native Patterson weighted by the sine term, since $\alpha_P - \alpha_{PH}$ is small if the F_H vector is small compared to F_{PH} and F_P vectors. The F_H^2 term is the heavy atom Patterson weighted by the cosine term, on average $cos^2(\alpha_{PH} - \alpha_H) = 0.5$. The $F_P F_H$ term produces a Patterson of the heavy atom-protein vectors, weighted by the sine term. The sign will also be changed at random by the cosine term, which will add to the noise on the Δ^2_{iso} Patterson. The final result is a Patterson map with peaks in the same place as the heavy atom Patterson but at half the peak height and with random noise. The heavy atom positions are located by looking at the Harker sections or lines, regions of the map have a higher than average density which arises because the vectors relating equivalent molecules (those atoms related by crystallographic symmetry) have one or two constant coordinates e.g. for spacegroup $P2_1$ there are two equivalent positions, the vectors relating these positions all lie on the Harker section at y=1/2. Complicated difference Patterson maps can be the result of having a high symmetry space group, extra noncrystallographic symmetry or a poor quality derivative. For complex difference Patterson maps there are several automatic Patterson search procedures available.

5.1.6 Direct methods

Direct methods derive the relative phases of the reflections by considering the relationships between the structure factors of the strong reflections. This approach is extensively used in small molecule crystallography it assumes:

- The electron density is always positive.
- The expressions are correct only for resolved and equal atoms.
- The relationships are good for only a few atoms.

Direct methods can be used to solve small protein structures providing that sufficiently high resolution data are available. The main application to protein crystallography is to find the heavy atom positions, since the number of heavy atoms is small. Problems arise when low resolution data give 'negative electron density' which is due to series termination effects. Another problem is that the heavy atoms sites have different occupancies, and cannot therefore be described as being equal. Direct methods use normalized structure factors. For proteins ΔF data are used. These data are reduced to a unique set and then normalized. There are two main problems with using direct methods (Sheldrick, 1991) on ΔF data, first, the negative quartets do not work since this depends on the identification of weak data, and the ΔF values do not identify weak reflections, only small differences. Secondly, it is not possible to estimate the probabilities for the triplet formula using ΔF data. However, direct methods do work for some proteins and it works best where there is translational symmetry, with a fixed origin and where there are no heavy atoms lying on special positions.

Difference Fourier synthesis

When estimates to the protein phases are available from other derivatives or from a molecular replacement solution, it is possible to use a cross-difference Fourier, to locate heavy atom sites. It is particularly useful for use with spacegroup $P2_1$, where the origin on the **b**-axis is not fixed until the first heavy atom position is found. This origin is then fixed and subsequent heavy atoms must be located with respect to it.

The Fourier synthesis has the coefficients:

$$m(F_{PH} - F_P)exp(i\alpha_P) \tag{5.11}$$

where m is the figure of merit, and α_P is the best protein phase, when SIR has been used to calculate the phase without any anomalous signal; the most probable phase is used when more than one derivative has been used to calculate the phase (Evans, 1991). This isomorphous difference Fourier is a highly distorted protein map with weak heavy atom peaks. It has been suggested that using the calculated phases on the heavy atom data alone should also locate the heavy atoms, and that the map should be less distorted (E.J. Dodson, personal communication).

The difference Fourier was used to check preliminary molecular replacement solutions. Phases calculated from the positioned search molecule were applied to the heavy atom difference data. If the difference Fourier peaks corresponded to peaks observed in the difference Pattersons it was seen as confirmation of the usefulness of the phases calculated from the MR solution.

5.1.7 Heavy atom refinement

Once the heavy atoms have been located the parameters that describe them are refined. All standard methods optimize the parameters that define the heavy atoms by minimizing the squared differences between the observed and calculated values. This type of refinement has poor discrimination and convergence because each amplitude term represents mostly noise and very little signal (see Section 5.1.5). All standard methods are statistically suspect (Bricogne, 1991) because all equations for the calculated values (F_H) depend on other observed quantities (F_{PH} or F_P). Most of these heavy atom refinement methods work best when a subset of the data, where the statistical bias is small, is used. There are two main types of refinement.

- Phaseless refinement: these methods use only the measured amplitudes, its main advantage lies in that it can be used when there is only one derivative available.

- Phased refinement: where phases are calculated from a previous derivative. Refinement can be carried in alternative cycles of phasing and heavy atom refinement or simultaneously.

Phasing and refinement are correlated, because we can minimize the phase error by either changing the phase (α_P) or by changing the parameters that define F_H. The heavy atom parameters that we are trying to refine are of two types, the global parameters and the local parameters. Global parameters include relative scale and temperature factors between the native and derivative data sets. The local parameters include occupancy, temperature factor and position of each heavy atom.

5.1.8 Phaseless refinement

This method uses the centric reflections only and works as long as there are more observations than there are parameters being refined (this is essential for all least squares refinement methods). As long as the selection of these reflections is uncorrelated with F_H, otherwise parameters, particularly occupancy, may be biased. When valid, centric refinement is perhaps the best method because it gives the best estimate of $\mathbf{F}_H$. It is best used for dihedral spacegroups, that is for spacegroups with two or more centric zones and when there are not too many heavy atom sites.

F_{HLE} refinement

This method uses acentric data and the anomalous signal in refinement, to improve the estimation of $\mathbf{F}_H$ (Dodson, 1976). F_{HLE} can be defined by:

$$F_{HLE}^2 = F_M^2 + F_P^2 - 2[F_M^2 F_P^2 - (\frac{K}{4})\Delta I^2]^{1/2} \tag{5.12}$$

Where

$$F_M = \frac{1}{2}||F_{PH}(+)|^2 + |F_{PH}(-)|^2| \tag{5.13}$$

$$\Delta I = |F_{PH}(+)|^2 - |F_{PH}(-)| \tag{5.14}$$

$$K = \frac{F'}{F''} \tag{5.15}$$

or

$$K = 2[\frac{< (F_{PH} - F_P) >}{< (|F_{PH}(+)|^2 - |F_{PH}(-)|^2) >]^{1/2}} \tag{5.16}$$

the residual minimized is:

$$R_2 = \sum_h w_h(|F_H|_{obs} - |F_H|_{calc})^2 \tag{5.17}$$

where

$$w_h = \frac{1}{Variance(|F_H|_{obs})} \tag{5.18}$$

The problems with this refinement method are that it is dominated by the centric reflections and it depends strongly on the weighting scheme. A good estimate of the anomalous signal is also needed.

'Heavy' refinement

This method is a modified version of the F_{HLE} refinement scheme (Terwilliger and Eisenberg, 1983). It uses a reciprocal space version of refinement with an origin removed Patterson. Removing the origin gives a higher value for the residual. When the origin is included the residual falls more rapidly and this means that it is difficult to remove incorrect sites and occupancies. The residual minimized is:

$$R = \sum_h w_h([(F_{PH} - F_P)^2 - < (F_{PH} - F_P)^2 >] - c[F^2_{H_{calc}} - < F^2_{H_{calc}} >])^2 \tag{5.19}$$

for acentric reflections $c = 1/2$ and for centric reflections $c = 1$. The centric and acentric data are treated separately.

Single isomorphous refinement has the disadvantage that no cross-phasing information is available from other derivatives and therefore, it is not possible to refine the origin in polar spacegroups. However, the method does enable incorrect heavy atom sites to be identified and eliminated. In addition there are no feedback errors due to common sites between two derivatives.

5.1.9 Phase refinement

Phase refinement involves alternative cycles of phasing and heavy atom refinement. A discussion of this technique requires that the calculation of the protein phases is described:

Phasing

Uncertainty in the isomorphous replacement experiment will introduce uncertainty in the resulting phases. There are four types of error in any isomorphous replacement experiment:

- Observational errors F_P and F_{PH} (random)
- Errors in scaling F_{PH} and F_P (systematic)
- Errors in heavy atom parameters (random and systematic)
- Lack of isomorphism

The propagation of errors through an experiment means that instead of discrete possibilities for the protein phase we get a probability distribution (Blow and Crick, 1959). The probability for a particular phase, α_P is given by:

$$P_j(\alpha) = Nexp[\frac{-x_j(\alpha)^2}{2E_j^2}] \tag{5.20}$$

where N is a normalization factor such that

$$\int_0^{2\pi} P(\alpha)d\alpha = 1 \tag{5.21}$$

The lack of closure error, x_j is given by:

$$x_j = \mathbf{F_{PH_{obs}}} - \mathbf{F_{PH_{calc}}} \tag{5.22}$$

and E_j is a measure of the total error:

$$< E >^2 = < \delta >^2 + < \epsilon >^2 \tag{5.23}$$

Where δ is the error in magnitudes of the measured reflections, F_{PH} and F_P and this is a 2D Gaussian function. It is estimated by measuring the same reflections from a different crystal or by comparison of the symmetry related reflections; it is the r.m.s. deviation of the magnitudes of the reflections. Errors in the heavy atom model and nonisomorphism, ϵ, have the form of a three-dimensional Gaussian. The error treatment for non-centrosymmetric reflections is complex, but the total error E_j can be estimated from the centric reflections where ϵ is well defined:

$$< E^2 >=< ([\mathbf{F_{PH}} \overset{+}{-} \mathbf{F_P}] - \mathbf{F_H})^2 > \qquad (5.24)$$

The error for a non-centrosymmetric zone becomes a convolution of ϵ and δ the result of which is a 'Gaussian ellipse'. A reasonable approximation to this function is the projection of the major axis of the ellipse onto F_{PH}. The total error of the reflections in the non-centrosymmetric case can be considered as residing in F_{PH} as long as F_H is small compared to F_{PH} and F_P.

A similar probability expression can be written for the anomalous scattering component. It is best to treat the anomalous and isomorphous data separately, since the errors associated with each are different. Anomalous data is intrinsically more accurate because there are no errors due to nonisomorphism, radiation damage or absorption. However the anomalous signal is smaller and therefore, the errors associated with counting statistics are larger.

As one would expect from the Harker construction probability functions calculated are bi-modal. Data from several derivatives are often needed so the phase probabilities from several derivatives are combined (Rossmann and Blow, 1961; Hendrickson and Lattmann, 1970). When combining the phase information from several derivatives the probability function becomes the product of the individual probabilities:

$$P(\alpha) = \Pi_j P_j(\alpha) = Nexp[-\sum_j (-\epsilon_j(\alpha)^2/2E_j^2)] \qquad (5.25)$$

Each time a new derivative is obtained the joint probability is recalculated. The Hendrickson and Lattmann formulation stores for each reflection the phase for

each derivative as a series with the coefficients A,B,C and D:

$$P(\alpha) \propto exp(Acos\alpha + Bsin\alpha + Ccos2\alpha + Dsin2\alpha) \tag{5.26}$$

The addition of new phase information only requires additions to these coefficients. This facilitates the combination of the different, or new, phase information. The most probable phase ($\alpha_{most\ probable}$), calculated from the bi-modal distribution, is used in phase refinement. The centroid of the distribution or the best phase (α_{best}) is used when calculating the 'best' Fourier map. Using the most probable phase in the latter case would give too much weight to some of the uncertain phases. The 'best' Fourier synthesis has the coefficients:

$$mF_P exp(i\alpha_{best}) \tag{5.27}$$

This gives the minimum mean square error in the electron density map. Where m is the 'figure of merit' or the cosine of the mean phase error. When the probability is sharp m is close to one; when the probability is flat, that is α_P can lie at angle, m is close to zero.

Phase refinement

Phase refinement minimizes the following residual:

$$R = \sum_h w_h(\mathbf{F_{PH}}_{obs} - \mathbf{F_{PH}}_{calc})^2 \tag{5.28}$$

or

$$R = \sum_h w_h(\mathbf{F_{PH}}_{obs} - k_{rel}\mathbf{F_P} + \mathbf{F_H})^2 \tag{5.29}$$

where

$$\mathbf{F_P} = |F_P|_{obs} e^{(i\alpha_{most\ probable})} \tag{5.30}$$

This procedure uses an estimate of the protein phases. With these phases held constant, the above residual is minimized with respect to the heavy atom parameters. New estimates for the protein phases are then calculated from these new heavy atom parameters, and the minimization is repeated. This procedure is

the same as refining the lack of closure (see Figure 5–3). Where the lack of closure is defined as:

$$\epsilon = \mathbf{F}_{\mathbf{PH}_{obs}} - \mathbf{F}_{\mathbf{PH}_{calc}} \tag{5.31}$$

where $\mathbf{F}_{\mathbf{PH}_{obs}}$ is the observed structure factor and $\mathbf{F}_{\mathbf{PH}_{calc}}$ is the structure factor calculated from the heavy atom model. The method's advantages lie in that it refines all the data jointly, that is the scale factors, relative occupancies and relative origins for several derivatives. A problem with this method is that incorrect heavy atom parameters feed bias into the phases which leads to serious errors if one derivative is wrong. This makes the refinement much slower to converge. By excluding the heavy atom parameters being refined from the phase calculation the rate of convergence is improved. If the refinement is carried out using the centric reflections only, or using only those reflections with a good figure of merit, refinement is quicker and has a larger radius of convergence. The correlation between the heavy atom parameters and the calculated phases can be accounted for using Bricogne's formulation (Bricogne, 1982).

An alternative refinement procedure (Sygusch, 1977) treats the phases as refineable parameters. The cosine and sine terms are refined with the heavy atom parameters. The increased number of parameters to observations means that this method needs to use a diagonal matrix approximation, thus the method loses some of the advantages it has gained. The major problem with this method is one of phase bias, a method for overcoming the problem of phase bias is by using maximum likelihood refinement (Otwinowski, 1991). This method applies the phase probability as a weight in the heavy atom refinement.

5.1.10 Maximum likelihood phase refinement

Phased refinement has recently been superseded by maximum likelihood refinement. There are two parts to any phase refinement: the refinement of heavy atom parameters and the phase calculation from these parameters. The former traditionally uses a least squares refinement, which considers only one phase for each reflection. This contrasts with the Blow and Crick treatment of

errors (Blow and Crick, 1956) which calculates a phase probability for each reflection and it is this probability that is used in phase calculation. The result of this treatment is that the heavy atom parameters can become severely biased due to the use of this single phase per reflection. MLPHARE, a heavy atom refinement program, introduces a likelihood function, based on the Blow and Crick estimate of errors, which is then used to estimate the heavy atom parameters. Here, the lack of closure residual is weighted by the probability of the phase being correct. Previously, this had been a problem in the refinement of single derivatives, where the modulus of the calculated heavy atom arrangement was fitted to the modulus of the observed differences. Therefore, any arrangement of heavy atoms could reduce the lack of closure and this could lead to wrong heavy atom sites being introduced which cannot be eliminated by using cross-phasing information from other derivatives.

5.2 Materials and methods

5.2.1 Preparation of heavy atom derivatives

The preparation of heavy atom derivative DADH crystals used heavy atom analogues to NAD^+ and direct binding of heavy atoms to the protein.

Heavy atom analogues to NAD^+

The binding of auricyanide ($Au(CN)_2^-$) and platicyanide ($Pt(CN)_4^{2-}$) to the NAD^+ binding site of LADH was used to prepare isomorphous derivatives in the determination of the structure (Eklund, *et al.*, 1975; Gunnarsson *et al.*, 1974). These derivative crystals were prepared by cocrystallizing soaking pre-existing DADH crystals with auricyanide and platicyanide anions.

Cocrystallization trials used both the hanging and sitting drop methods. Crystallization was as described in chapter 2, except excess heavy atom complex, 10^{-3}M $KAu(CN)_2$ or 10^{-3}M $K_2Pt(CN)_4.3H_2O$ was added to all crystallization

buffers. Trials were carried out with a pH range 6.9 - 7.1, and with varying concentrations of PEG 4000 (6 - 20%). Cocrystallization trials with fresh protein, produced form A crystals in the presence of both $KAu(CN)_2$ and $K_2Pt(CN)_4.3H_2O$. These crystals were badly twinned and not suitable for X-ray diffraction studies.

Test crystals were soaked in artificial mother liquor, 50 mM citrate phosphate buffer at pH 7.0 with 0.2% sodium azide, 10^{-4}M DTT and 20-22% PEG 4000. To this mother liquor 1-10 mM $KAu(CN)_2$ or $K_2Pt(CN)_4.3H_2O$ was added and the crystals were observed for up to one week. As little as 1 mM $KAu(CN)_2$ caused small test crystals to crack after one day. A larger crystal (0.2 x 0.2 x 0.2), which had been soaked in a 10 mM $KPt(CN)_4$ solution for 20 hours, diffracted X-rays to better than 3 Å but showed a 10% change in the dimension of the **c** axis. Data were collected on a crystal soaked in 5 mM $KPt(CN)_4$ for 20 hours, however this showed no substitution (see Table 5–4).

Direct binding of heavy atoms

Since the supply of good protein crystals was poor, soaking experiments were first carried out on small test crystals:

- in the dark
- in freshly made heavy atom solution
- at 10°
- in 1 mM solutions
- observed for one week

Promising soaking experiments were then repeated on larger crystals and when possible, data were collected. The DADH crystals were grown in citrate phosphate buffer so it was noted that uranyl and lanthanide compounds form complexes with the phosphate.

DADH has two free cysteines per monomer. These groups react to form covalent bonds with mercury and this was the starting point for the preparation of isomorphous derivatives of the DADH form B crystals. When observing the reactivity of sulphydryls, the presence or absence of reducing agents needs to be taken into consideration. The covalent reaction of sulphydryls is sometimes very quick, in the absence of reducing agent, e.g. a 0.01 mM solution gave full occupancy after a one hour soak in the case of iron superoxide dismutase (Ringe, 1983). However, by adding reducing agent to a soaking solution it is possible to control the rate of the heavy atom binding since the reducing agent prevents heavy atom binding to the cysteines (Katz *et al.*, 1985).

The DADH form B crystals were grown in the presence of DTT. Since the presence of DTT had been necessary for the stability of the crystals, initial heavy atom soaking trials targeting the cysteines, were carried out in the presence of 10^{-4}M DTT. The heavy atom complex, usually a mercury complex, was present in excess, concentrations of at least 1 mM being used. This approach yielded weakly substituted derivatives. In later trials DTT was removed from the crystals before heavy atom solutions were added.

A summary of the soaking experiments carried out is shown in Table 5–1.

Mercury soaks: Initial soaks with mercury chloride, mercury acetate, mercuric orthophosphate, p-chloromercuri benzene sulphonate (PCMBS) on small crystals showed that several mercury compounds were suitable for data collection. Data were collected on a crystal which had been soaked in 1 mM 2-chloromercuri-4-nitrophenol (CMN) for 20 hours at 10° but was found not to have a substituted heavy atom. This was probably due to the presence of DTT in the buffers. Further trials were carried out in the absence of DTT. DTT was removed from the protein crystals by backsoaking the crystals in artificial mother liquor, with a slightly increased precipitant concentration. The mother liquor had been purged with nitrogen gas. Crystals were backsoaked for 3 hours and then soaked in heavy atom solution. Finally, the crystals were backsoaked again in nitrogen purged buffer to remove excess, unreacted heavy atom. Three derivatives were prepared in this way: a crystal soaked for 12 hours in 6 mM

Compounds	Conc. (mM)	Time	Temp. (C)	Comments
Mercury Compounds				
Mercury Acetate	3	24		Did not diffract X-rays
	1			?
$HgCl_4$	3	12 hours	10°	Incomplete data set collected
	3	20 hours	10°	DDT removed. Old crystal
DMA	2			Low solubility in buffer
Bakers mercurial	3	1 week	10°	Crystal survived several days
Mercurial orange	2	1 day	10°	Crystal dissolved
PCMBS	1	1 week	10°	Crystal started to crack after 5 days
Mercuric orthophosphate	2	1 week	10°	Crystal cracked after 5 days
Ethyl mercury phosphate				Failed
Mersaly				Failed
MMA	1	1 week	10°	Crystal survived
2-chloromercuri-4-nitrophenol	3	12 hours	10°	Data collected to 3 Å
	6	12 hours	10°	Diffraction to 5Å
$HgI_2.KI.1\frac{1}{2}H_2O$	6	?	10°	Failed
	2	20 hours	15°	Crystals did not diffract
Gold Compounds				
$KAu(CN)_2$	4		10°	Crystal broke up
	1	24 hours	10°	Weak diffraction
	1	1 week	10°	Cracked after several days
$NaAuCl_4.2H_2O$	4	2 days		Crystal survived
	1	1 day	10°	Crystal survived
	1	6 hours	?	Diffracted to low resolution
$KAuCl_4.2H_2O$	0.1	1 day	?	Crystal destroyed
$KAuI_4$	1	6 hours	?	No diffraction
Platinium Compounds				
K_2PtCl_4	1	6 hours	10°	The crystal was destroyed
	0.1	20 hours	4°	Data collected. Weak derivative
	0.5	20 hours	10°	Data collected
H_2PtCl_6	5	19 hours	10	Cell dimensions changed
$K_2Pt(CN_4).3H_2O$	5	20 hours	10	No substitution
$Pt(NH_3)_4Cl_2.H_2O$	1	24 hours		Crystal cracked
	1	6 hours		Crystal ok.
	1	20 hours	10	Data collected.
	5	20 hours	10	Cell changed. Poor data
	10	12 hours	4°	Diffraction died rapidly in X-ray beam
$Pt(NH_3)_2Cl_2$	1	24 hours	10°	Crystal lost
(cis-Pt)	1	6 hours	?	Diffracts to 4Å
$Pt(NO_4)$	3	12 hours	10°	Did not diffract.
Miscellaneous				
$CdBr_2.4H_2O$	10		10°	Crystal survived.
Na_2IrCl_6				Not isomorphous.
	1.5	6 hours	4°	Not isomorphous.
	0.1	24 hours	10°	Not isomorphous.
	1	1 hour	20°	Crystal disordered. Try again.
K_2PdCl_4	1	3 hours		No diffraction.
K_2OsO_2	1	7 hours		Diffracted to low resolution.

Table 5–1: Summary of all soaking experiments carried out on DADH form B crystals

2-chloromercuri-4-nitrophenol diffracted X-rays to 5Å; A crystal soaked in 3 mM 2-chloromercuri-4-nitrophenol (GCMN) for 12 hours diffracted X-rays to better than 3 Å; and a crystal soaked for 20 hours at 10°C in 1mM mercury chloride solution (GHG). Details of the data collected on these derivative crystals is discussed later (see Section 5.2.2).

The removal of DTT seemed to disrupt the order of the crystal, surface cracks appeared and the crystals diffracted poorly.

Platinum soaks: Platinum chloride is a popular heavy atom complex for preparing isomorphous derivatives. Soaking small test crystals in 1 mM K_2PtCl_4 destroyed the crystals. However, soaking in 0.1 mM and 0.5 mM K_2PtCl_4 (GPTC) gave weakly substituted derivative crystals. The crystal soaked in 0.1 mM K_2PtCl_4 (PTC) was the most useful derivative. Data were collected on a crystal that had been soaked for 20 hours in 1 mM tetrammine platinum (II) chloride solution (TAPC): there was no substitution but a crystal soaked in 5 mM TAPC under the same conditions was non-isomorphous to the native crystal.

Gold soaks: 1 mM $NaAuCl_4$ disrupted the crystal after only 6 hours soaking. Future soaks should aim to reduce this disruption by using lower concentrations of heavy atom solution for shorter time periods.

Iridium soaks: Soaks with 2.5 mM $NaIrCl_6$ for 6 hours changed the cell dimensions by 6% . The concentration of iridium was reduced to 1 mM but significant changes in the cell dimensions were still observed.

5.2.2 Data collection

Data were collected on a number of soaked crystals using a Xentronics detector and processed using XDS (Kabsch, 1988). The large cell was used to integrate the data. Details of the data collection and processing are found in chapter 3. All data sets were internally scaled using the CCP4 programs ROTAVATA and AGROVATA (LCF versions) and the results are summarized in table 5–2. ROTAVATA and AGROVATA scale data internally and analyze the standard

Resolution (Å)	R_{sym}(%) [Completeness (%)]							
	PTC		CMN		GCMN		GHG	
8.81	2.5	[98.1]	2.3	[98.1]	3.7	[93.9]	1.5	[93.8]
6.32	3.2	[97.1]	2.3	[97.9]	7.3	[92.5]	3.3	[96.6]
4.00	3.2	[97.0]	2.4	[98.8]	11.2	[79.9]	4.8	[94.9]
2.83	7.3	[76.5]	7.0	[93.9]	23.3	[53.8]	4.7	[80.0]
2.48	5.6	[19.1]	20.7	[38.4]				
Overall R_{sym}	4.0		2.8		10.1		2.2	
No. refs	35955		26008		14516		16331	

Table 5–2: Crystallographic statistics as a function of resolution, for heavy atom soaked crystals.

Data set	GHG	PTC	TAPC	CMN	GCMN	PTCN
Compound	$HgCl_2$	$KPtCl_4$	$Pt(NH_3)_4Cl_2.H_2O$	$HgClC_6H_2(NO_2)OH$		$KPt_2(CN)_4$
Conc (mM)	1	0.1	1	1	3	5
Soak(hours)	20	20	20	20	20	20
Resolution (Å)	3.15	2.88	2.88	2.67	3.15	2.88
Unique refs.	7783	9996	10187	13319	6700	9746
Completeness (%)	90.0	88.8	90.5	94.5	78.0	86.6
Redundancy	1.5	1.1	1.6	1.5	1.4	1.5
$< I/\sigma I >$	24.1	11.8	10.1	21.4	6.5	8.9

Table 5–3: Summary of crystal preparation and data collected on heavy atom soaked crystals. All data has been reindexed to small cell and the completeness calculated.

deviations estimated by the integration program. Adjustments are then made so that the standard deviations have a normal distribution with a mean of zero and a standard deviation of one. The data were then reindexed to the small cell. Native and derivative data were scaled together using SCALEIT and analyses were carried out. The preparation conditions for these crystals are summarized in Table 5–3.

The analysis carried out in SCALEIT calculated the fractional mean isomorphous difference (R_{iso}), the mean isomorphous difference (D_{iso}) and the normal probablity plots. Table 5–4 summarizes these analysis for the data collected on heavy atom soaked crystals. The preliminary analysis of the data from the heavy atom soaked crystals shows that there has been some heavy atom substitution in the GHG, PTC, GCMN and possibly the GPTC data sets but these changes are small. These data were isomorphous with the native (see Figure 5–6 and Figure 5–7).

The analysis of the normal probability distribution for the data collected on the crystals soaked in heavy atom solution indicate that there is a systematic

Data set	GHG	PTC	GPTC	TAPC	CMN	GCMN	PTCN
No. refs	7880	9902	14131	9923	8849	6048	9472
$<D_{iso}>$	1144.9	1033.7	633.6	835.2	384.6	2566.6	740.2
ND Gradient	11.14	7.87	6.44	5.15	4.85	14.46	5.64
R_{iso}	14.2	13.6	11.8	5.3	6.4	16.9	6.5

Table 5–4: Table summarizing the results of the analysis of the heavy atom soaked data sets. The Normal distribution plot should have a gradient of 1.0, if there is no systematic change between the native and derivative data sets; it is suggested that substitution of heavy atoms gives a gradient of 15.0. $R_{iso} = \frac{|F_{PH}|-|F_P|}{|F_{PH}F_P}$.

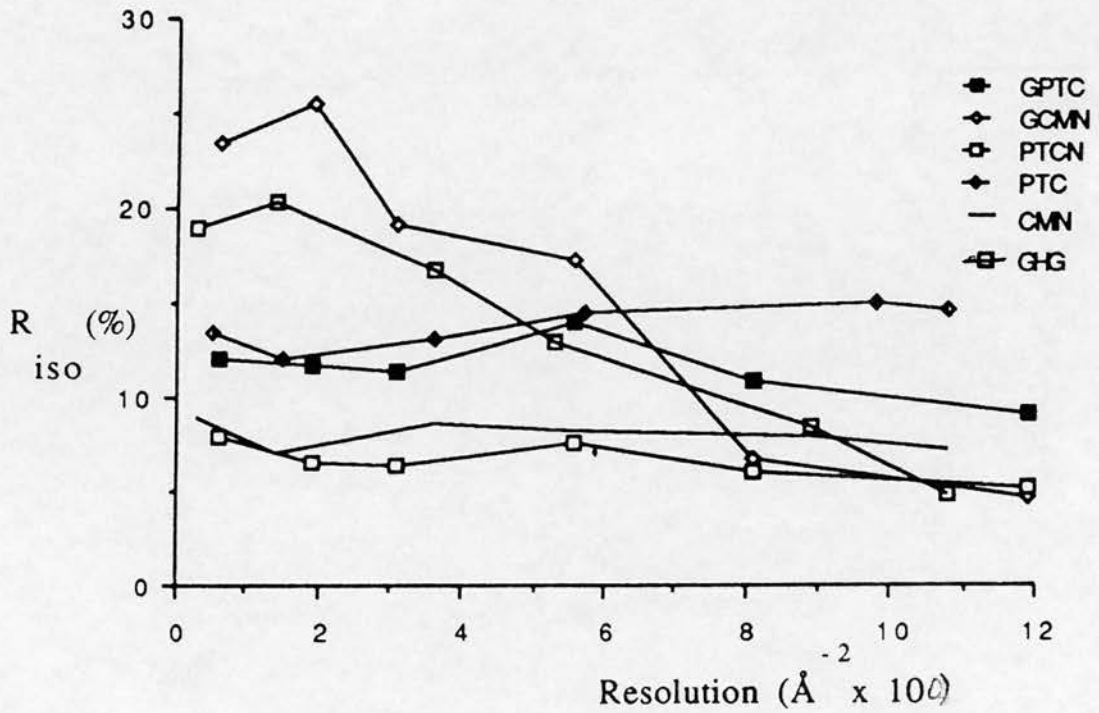


Figure 5–6: Plot of R_{iso} as a function of resolution.

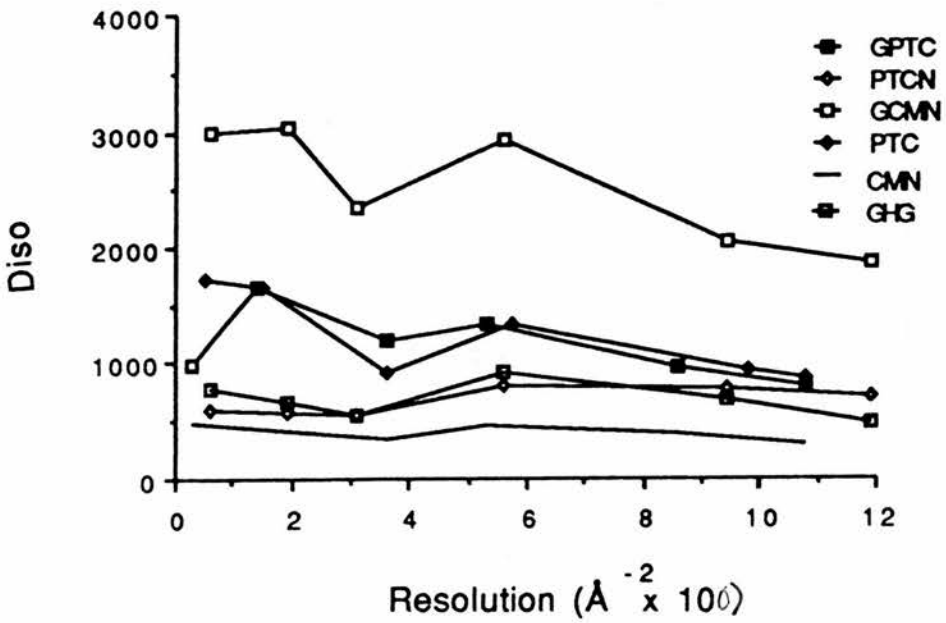


Figure 5–7: Plot of D_{iso} as a function of resolution.

Data set	GHG	PTC	GCMN
No. refs.	7886	11188	6072
$<D_{iso}>$	1142.9	1053.6	2397.7
Resolution	2.8	2.5	3.1
ND Gradient	10.65	7.98	13.06
R_{iso}	17.4	18.1	18.3

Table 5–5: Preliminary analysis of the differences between the data collected on heavy atom soaked crystals, and the CMN data set. These results can be compared directly with those in table 5–4.

difference between the native and the derivative data, but that this change is small.

GHG, PTC and the GCMN are possible derivatives since they show significant changes when compared to the native data set. Analysis of these changes as a function of resolution indicates that these changes decrease with increased resolution i.e. the changes are not due to nonisomorphism.

Problems encountered with the native data set during the molecular replacement study, where cross rotation function results were incompatible with calculations carried out on other data sets (see Section 4.2.3), indicated that there were some inconsistencies within the native data set. To test this, a control experiment was carried out by substituting the CMN data as an alternative native data set (since the $R_{iso} = 6.4\%$ between the native and the CMN data indicated that there had been little or no heavy atom substitution). A summary of the preliminary analysis of the GCMN, PTC and GHG data sets against the CMN data are shown in table 5–5. The Harker sections ($v = 1/2$) of the difference Patterson maps calculated using the CMN data are shown in figure 5–11, figure 5–12 and figure 5–13.

Location of heavy atom positions

The heavy atom positions are found using difference Patterson maps and difference Fourier synthesis. A difference Patterson map is a good indicator of derivative quality. A clean map indicates an isomorphous derivative with few or single sites of substitution. The difference Fourier synthesis was computed using

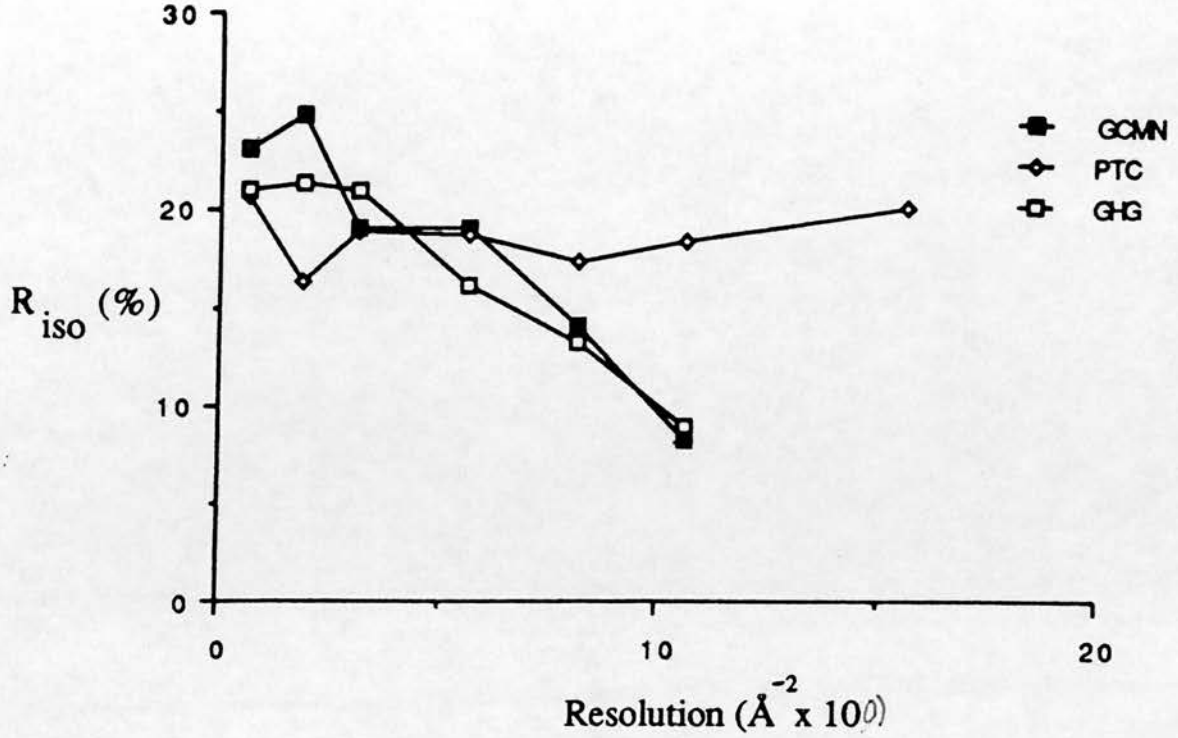


Figure 5–8: Plot of R_{iso} as a function of resolution. For difference data using CMN data set as the native data set.

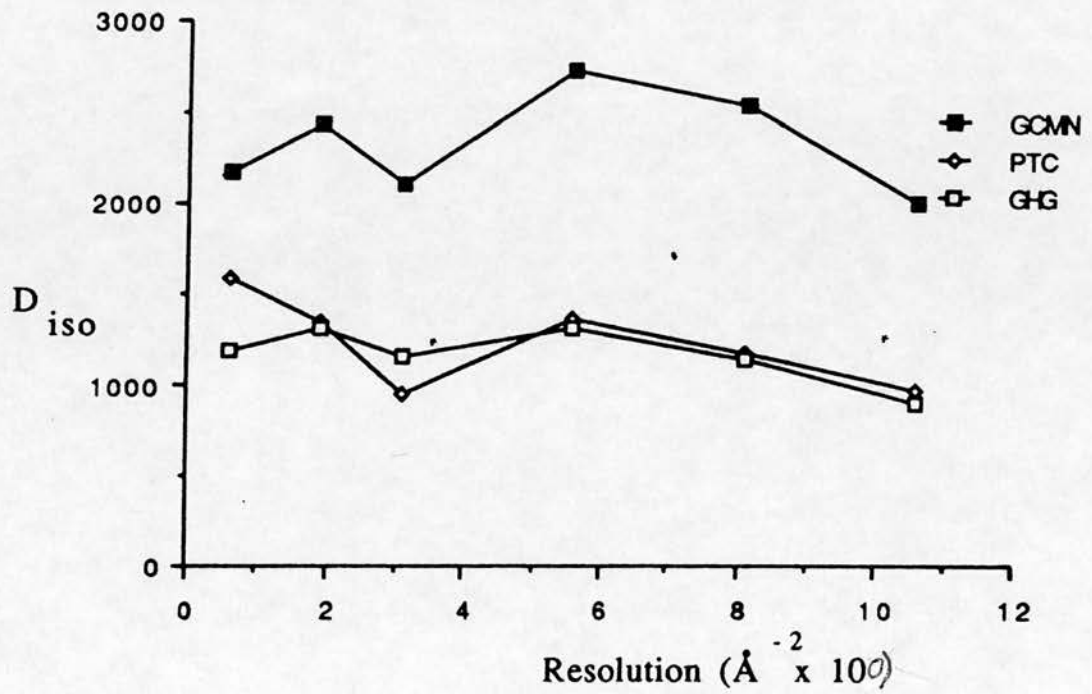


Figure 5–9: Plot of D_{iso} as a function of resolution. Using CMN data set to replace native data set.

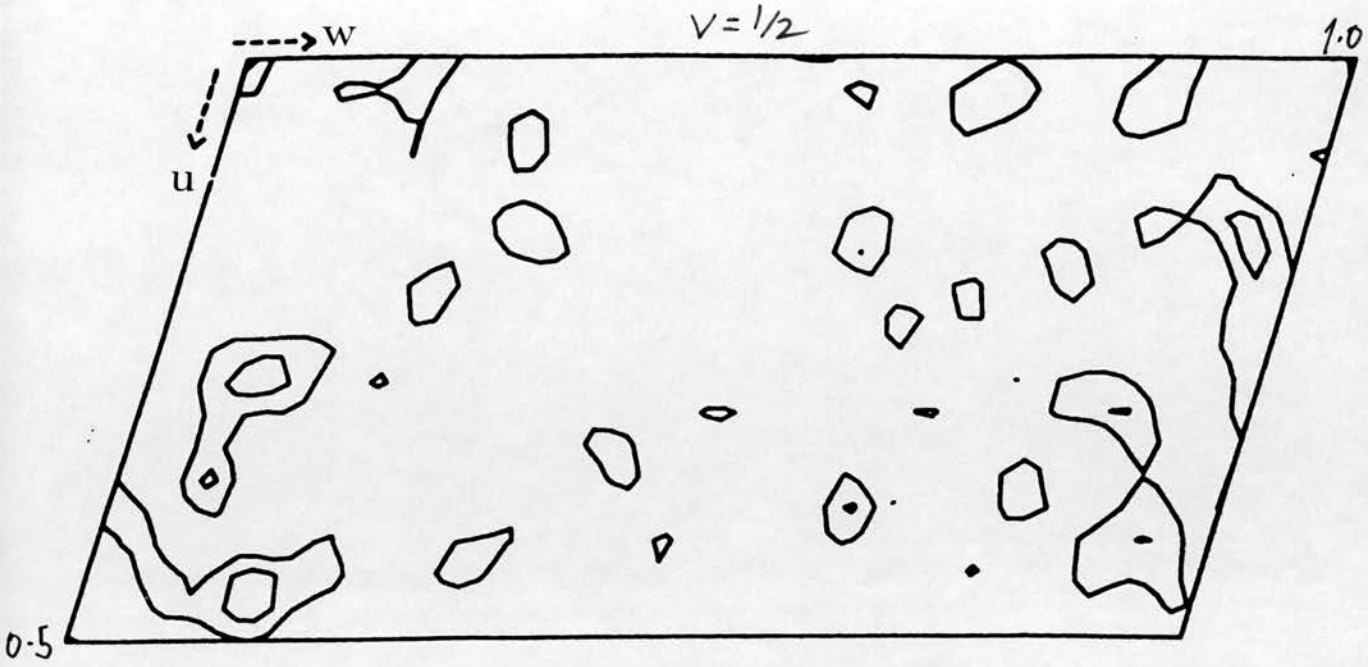


Figure 5–10: Harker section for CMN *versus* native Patterson.

the phases obtained from the molecular replacement solution. An accurate molecular replacement solution gives a difference Fourier map which correlates with the difference Patterson map.

The difference Pattersons were recalculated using the CMN data set as a native data set. Preliminary analysis between the data from the heavy atom soaked crystals, and the CMN data are summarized in table 5–5. The Harker section for the CMN *versus* native difference Patterson is shown in figure 5–10.

When preparing the difference Patterson map, several precautions were taken to give a good map:

- Rejection of outliers, since the Patterson function is dominated by a few large terms, therefore large rogue reflections should be excluded from the Difference Patterson synthesis e.g. reject reflections 3 x mean isomorphous difference. SCALEIT outputs a suitable cutoff level.

- Compare difference Pattersons calculated for independent ranges of resolution, for example 10-7 Å and 7-4 Å. Significant peaks should be present in both ranges.

- Use artificially sharpened structure factors by restricted range of data.

- The solutions were tested by plotting the Pattersons for promising solutions.

The Harker section for spacegroup $P2_1$ lies at y=1/2. Each section covers the map from x=0 - 0.5 and z=0 - 1.0. The plots were contoured at the r.m.s. deviation of the maps. The Harker sections for the difference Pattersons for the GHG, PTC and the GCMN data (see Figures 5–11, 5–12, 5–13) are noisy, possibly indicating poor data or multiple heavy atom sites with low occupancies.

The difference Patterson calculated using the CMN data in place of the native data seemed to be less noisy than the maps calculated using the native data sets. There were cases where weak peaks in the heavy atom *versus* native map were stronger in the heavy atom *versus* CMN map and *vice versa.* However, generally the use of the CMN data seems to improve the quality of the plots. The possible exception to this is the PTC difference Patterson map. This map became more noisy and the peaks appeared streaked. Both the GHG and GCMN difference Patterson have a strong peak at $u = 0$, $v = 1/2$, $w = 0$. This peak was refined for the GCMN data and gave a low occupancy and a negative temperature factor indicating a wrong site.

Direct methods were run routinely on all of the data sets but failed to resolve the heavy atom positions satisfactorily.

Difference Fourier syntheses

Heavy atom positions, located using difference Fourier synthesis using phases derived from the molecular replacement solution, were checked against peaks found in the difference Patterson maps and against cross vector peaks.

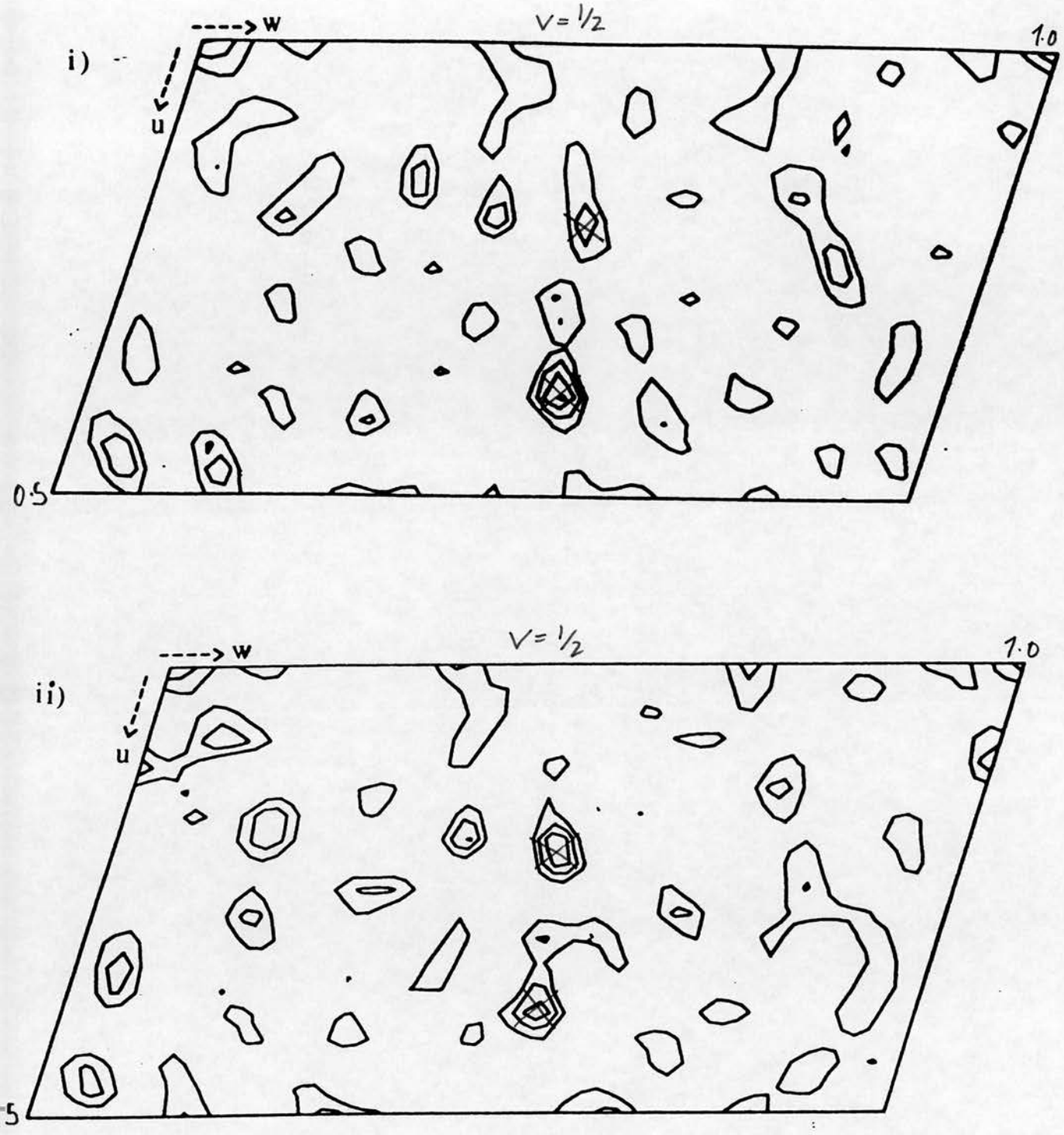


Figure 5–11: Harker section for difference Patterson map for (i) the GHG and native differences and (ii) the GHG and CMN differences. Contours at r.m.s. deviation of the map.

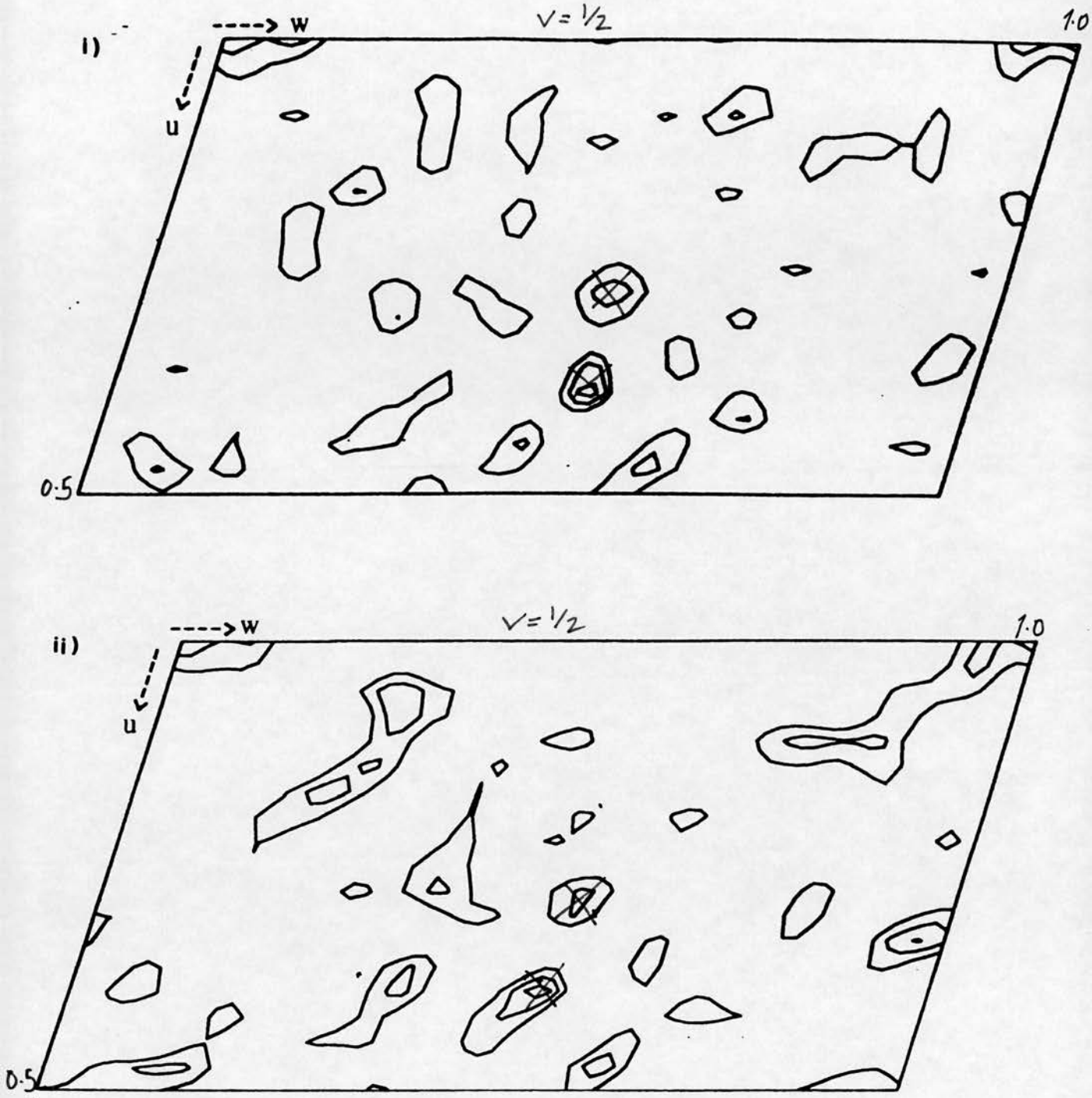


Figure 5–12: Harker section for the difference Patterson for (i) the PTC and native differences and (ii) the PTC and CMN differences. Contours at r.m.s. deviation of the map.

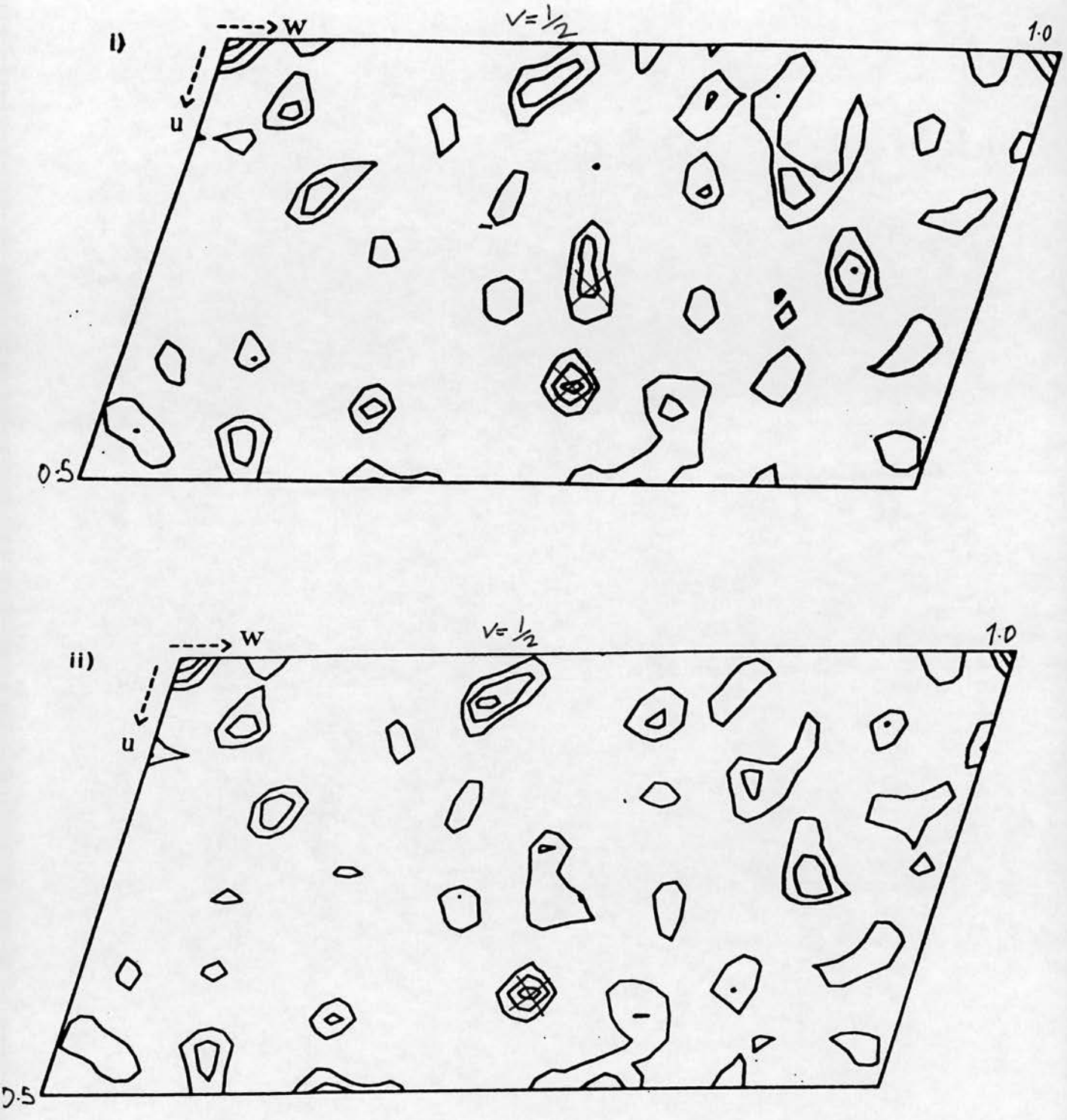


Figure 5–13: Harker section for the difference Patterson for (i) the GCMN and native differences and (ii) the GCMN and CMN differences. Contours at r.m.s. deviation of the map.

Data set	Heavy atom position			Occupancy	Anomalous occupancy	B-factor
GHG	0.093	0.393	0.240	1.280	2.304	26.345
	0.351	0.295	0.291	6.356	0.731	42.240
	0.573	0.069	0.387	12.549	0.802	4.155
	0.536	0.120	0.936	0.755	-1.316	28.089
	0.418	0.364	-0.026	0.234	0.273	-80.795
GCMN	0.573	0.069	0.385	27.375	11.988	-0.709
	0.352	0.296	0.291	14.415	2.260	40.318
	0.399	0.479	0.525	1.044	-1.689	-33.813

Table 5–6: Refinement statistics for heavy atom positions

	MIR parameters	
	GCMN	GHG
No. refs	4684	5363
Total phasing power	1.4	1.3
Total Rcullis	70	73
< FOM > (acentric)	0.430	

Table 5–7: Heavy atom refinement statistics.

5.2.3 Heavy atom refinement

Heavy atom positions were refined using the maximum likelihood phase refinement (MLPHARE). The statistics for the refinement of heavy atoms are shown in tables 5–6 and table 5–7.

A good derivative should have a large phasing power i.e. it should have a large isomorphous change and a small lack of closure (see figure 5–8 and figure 5–14).

The most useful derivatives were the two mercury derivatives. The heavy atom positions located for the PTC data did not refine, so were discarded. As expected the GHG derivative had been substituted at more sites than the GCMN, derivative. The 2-chloromercuri-4-nitrophenol complex is larger and more likely to be prevented from binding by steric constraints. The two major sites in both derivatives are common sites; x = 0.57, y = 0.07, z = 0.39 and x = 0.35, y = 0.30, z = 0.29. The occupancies of these sites do not seem to indicate that they are equivalent sites related by noncrystallographic symmetry. Although binding at one site might be restricted due to the arrangement of the molecular packing in the crystal.

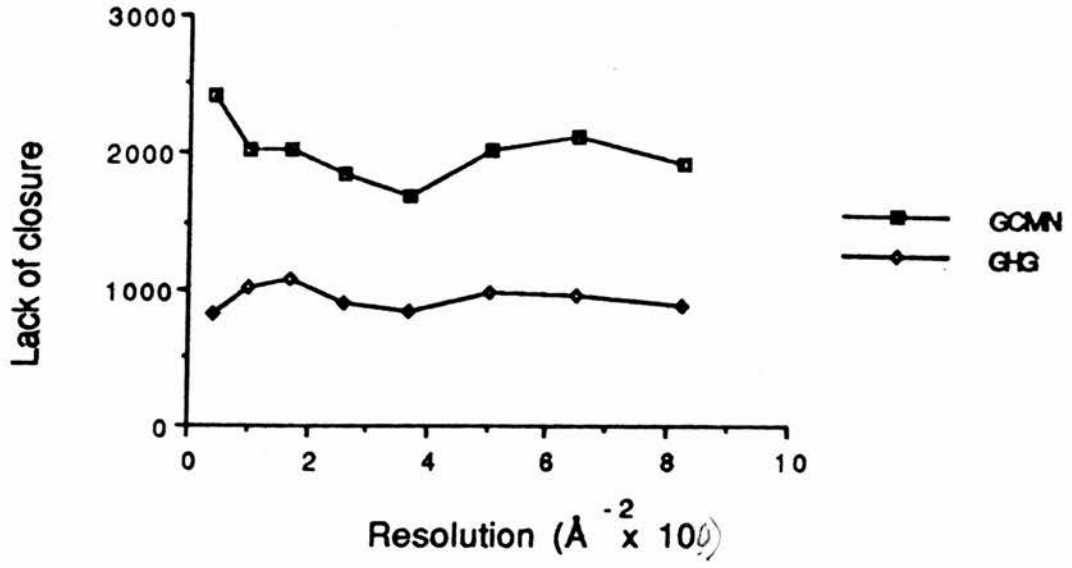


Figure 5–14: Lack of closure (acentric zone) as a function of resolution for GCMN, GHG derivatives. Where the lack of closure is $|F_{PH_{obs}} - F_{PH_{calc}}|$.

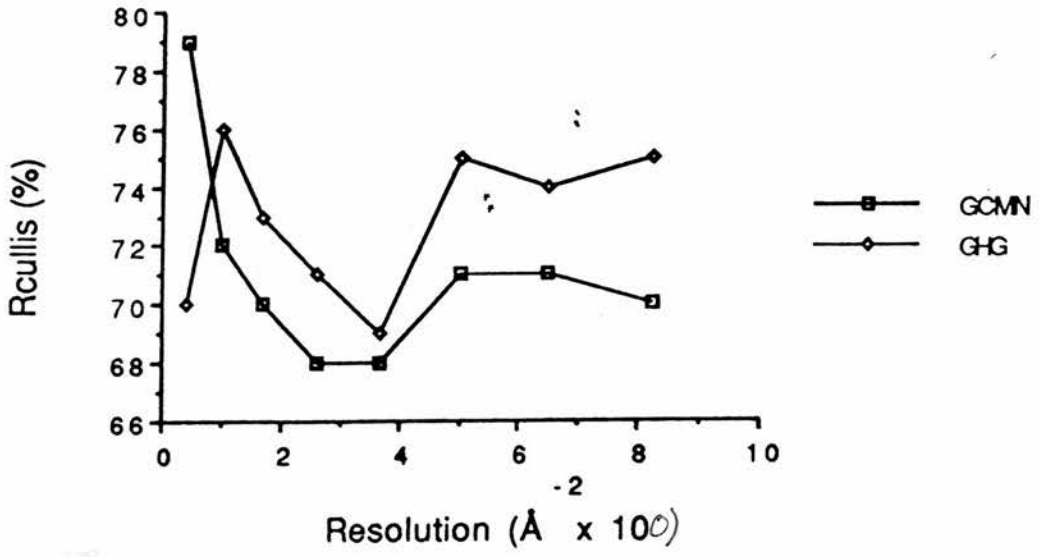


Figure 5–15: Rcullis as a function of resolution for GCMN, GHG derivatives. Where Rcullis is defined as $\frac{\Sigma||F_{PH}+/-F_P|-F_{Hcalc}|}{\Sigma|F_{PH}-F_P|}$

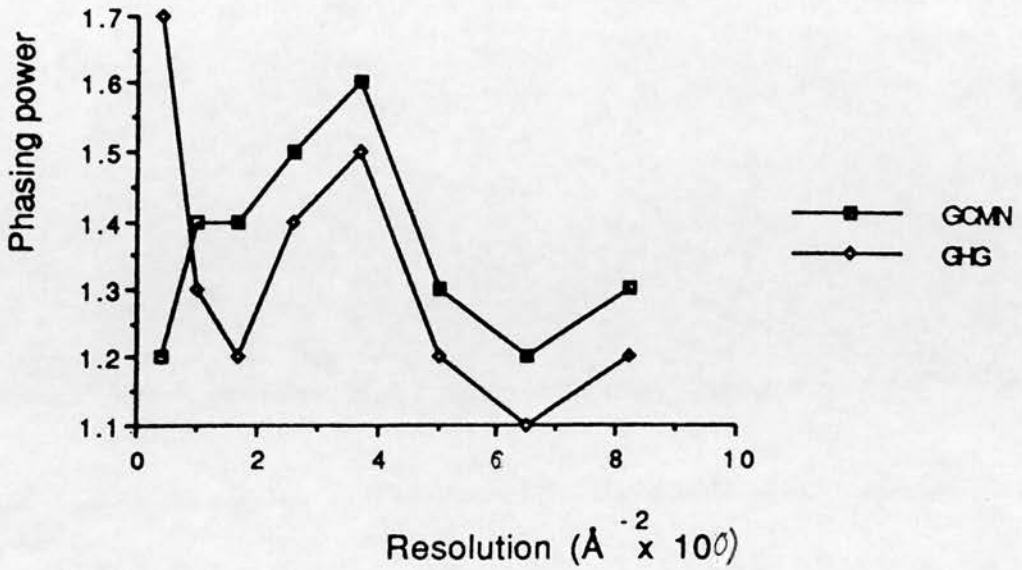


Figure 5–16: Phasing power as a function of resolution for GCMN, GHG derivatives. Phasing power F_{Hcalc}/lack of closure

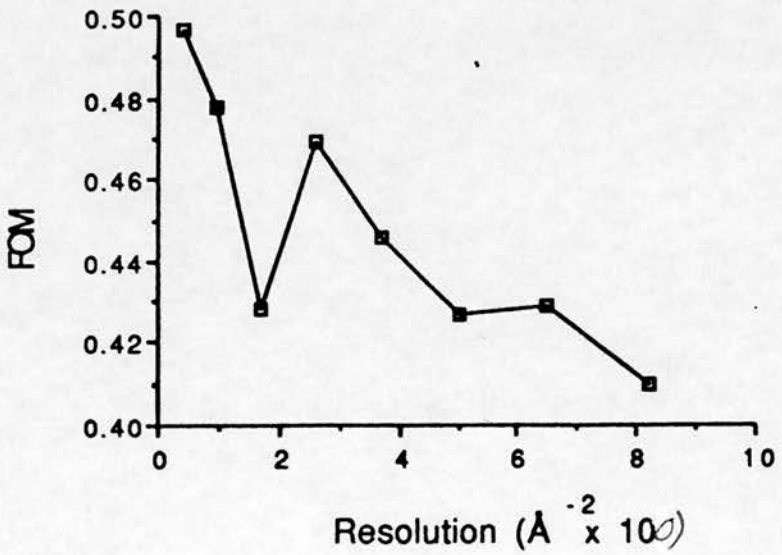


Figure 5–17: Figure of merit (FOM) as a function of resolution.

5.2.4 Electron density map

The heavy atom derivatives are weakly substituted and the phasing statistics are correspondingly poor. An electron density map was calculated with data between 20 - 3.5 Å using the isomorphous replacement phases. Since the major sites in both cases are common sites, the phasing is essentially one of single isomorphous replacement. Although some features of the map were apparent the map was too noisy to allow for the identification of the solvent boundary. It was not possible to identify any noncrystallographic symmetry in the density. This MIR map was subjected to map improvement techniques (see Chapter 6).

5.3 Discussion

It seems that DADH crystals are very susceptible to damage by the heavy atom solutions. Addition of heavy atoms to the mother liquor, in some cases causes surface cracks to appear on the crystal, these crystals diffract well but have a high mosaic spread. Removal of DTT from crystals has an obvious effect on the surface of the crystal. But removal of DTT is essential if mercury derivatives are to be obtained.

The addition of low concentrations of heavy atom solution seem to result in reasonable isomorphous changes (i.e. R_{iso} = 15 - 25%) and these changes do not seem to be due to nonisomorphism (see plots of R_{iso} and D_{iso} as a function of resolution).

The Harker sections of the difference Patterson maps appear to be noisy. This is probably due to a combination of multiple sites of substitution and the presence of noncrystallographic symmetry. Although the presence of noncrystallographically related heavy atom sites is not clear at this stage.

5.3.1 Future work

It is clear that improvement of the MIR map requires more derivative data. Which will require that more heavy atom soaked crystals are screened. There may be some advantage in recollecting the GCMN and GHG data with some variations in the soaking concentrations so that heavy atom occupancies are obtained e.g. work at a system that reduces osmotic shock and damage by removal of DTT from the crystals. It is advisable that a systematic study on the effect of removing DTT from the crystals and how this affects the diffraction from these crystals. However, such an approach requires many crystals and available data collection facilities.

Chapter 6

Map Improvement

6.1 Introduction

When the initial phases obtained from MR or MIR experiments are not good enough to produce an interpretable electron density map, it is possible to refine these phases by applying restraints which have been determined by the known or expected physical properties of the crystal (Podjarny and Rees, 1991). The 'refinement' process involves density modification in real space, back transformation of the modified map and then combining these 'modified' phases with the initial MIR phases (see Figure 6–1).

There are various methods for calculating these physical restraints and applying them:

- Noncrystallographic symmetry: the presence of noncrystallographic symmetry leads to a redundancy in information for the asymmetric unit. With accurate knowledge of the position of symmetry elements that relate equivalent molecules or subunits within the asymmetric unit, the maps covering these equivalent molecules can be averaged. The averaged map is then back transformed and the resulting phases are combined with the initial MIR phases. The greater the degree of noncrystallographic symmetry, the greater the power of the method.

- Real space density modification: this covers a range of methods, the most common of which, is solvent flattening. The solvent flattening procedure makes two assumptions: that the density in solvent regions is featureless and that the protein region should have no negative densities (Wang, 1985). The technique can also be carried out in reciprocal space (Leslie, 1988). Histogram matching is another method of density modification (Zhang and Main, 1988). The method involves fitting the density histogram of an observed image to the histogram of a 'good' image. The observed image can then be improved in a systematic way until the

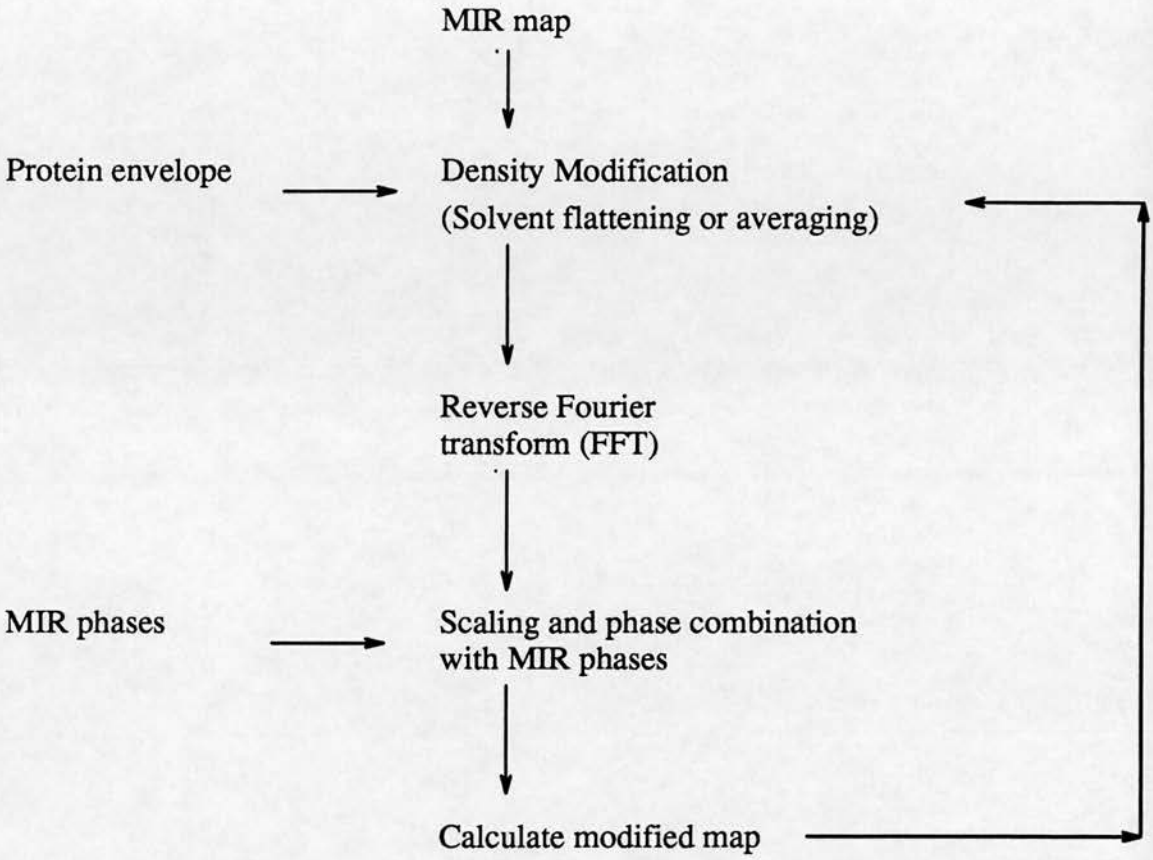


Figure 6–1: Flowchart showing the steps involved in a cycle of DM

histogram matches the good image histogram. Phases are then calculated from the modified map and combined as above.

- Direct methods: this uses a theoretical relationship between the relative phases of the structure factors of the strong reflections. The relationship assumes that the electron density is always positive and the scattering 'randomly' arrayed. This is a reciprocal space method of phase improvement.

In some cases it has been found that using the phases calculated from the solvent flattened map without combining with initial phases works best, however the quality of the starting phases in this case has to be good (Vellieux *et al.*, 1988).

When merging the phases the density modification process converges when the phase difference between the initial MIR phases and the density modified phases reaches an arbitrary minimum. This phase convergence means that the starting phases have incorporated the density modification constraints but does not imply that the phases are correct and accurate.

6.1.1 Phase extension

Phase extension methods use the known phase set to extend slowly the phases to higher resolution. The aim of this method is to phase all the native reflection data available. The success of this procedure depends on the goodness of the starting phases. In solvent flattening and molecular averaging when the known phase set is large with respect to the unknown phase set, phase extension can be used.

6.2 Methods

Density modification and direct methods were carried out using SQUASH (Zhang and Main, 1988; Cowtan, 1991). This program can apply solvent flattening, histogram matching, Sayre's equations and noncrystallographic symmetry averaging. The effect of these modification processes depends on the quality and resolution of the available data. Solvent flattening is effective with data of all resolutions, but is less efficient with crystals where solvent content is low.

The density histogram (a plot of the probability of an electron density against observed electron density), of a protein structure is independent of protein structure. So histogram matching can be used in the structure determination of an unknown structure . The density histogram is dependent on resolution and overall temperature factor and the technique becomes less effective at resolutions of less than 4 Å. Sayre's equations work if there are good starting phases to at least 4 Å, but preferably nearer 2 Å resolution. Both the histogram matching

and Sayre's equations depend on absolute electron density, therefore the absolute temperature factor for the data has to be determined prior to running SQUASH. An estimate of the absolute temperature factor can be calculated using the CCP4 program WILSON.

6.2.1 Solvent flattening

The solvent content of a crystal is known from preliminary X-ray diffraction studies of the native crystals (Matthews, 1968). Generally a conservative estimate of the solvent content is used in solvent flattening procedures since this reduces the risk of segmentation of the protein envelope. DlADH crystals have a solvent content of 44% but solvent flattening procedures used 35% solvent content. An envelope for the protein density is determined automatically by looking at the weighted electron density map (100 - 3.5 Å) calculated from MIR phases or from MR phases (Wang, 1985; Leslie, 1988) and protein regions and solvent regions are distinguished. This envelope is imposed upon the original map and the region outside the envelope, which is solvent, is levelled. For noncrystallographic symmetry averaging the mask has an additional function: that of identifying separate protein subunits. The molecular subunits are usually identified by looking at a low resolution electron density map. For both solvent flattening and noncrystallographic symmetry averaging the envelope regions should make physical sense e.g. symmetry related envelopes should not inter-penetrate each other. The envelope was re-determined after every 4 cycles of density modification.

6.2.2 Convergence of density modification

The course of density modification is monitored by looking at the phase differences between the current cycle and the previous cycle, these should decrease and stabilize as convergence is reached. The phase differences for the high resolution data should be noted since these converge more slowly than the low resolution data.

No. of cycles	12
Resolution range of maps	100 - 3.5 Å
No. of reflections	6205
R cycle 1	54.4 %
< FOM >	0.45
Correlation between F_{obs} and F_{calc}	0.56
Correlation between old and new maps	0.79
R final	60.3 %
< FOM >	0.79
Correlation between F_{obs} and F_{calc}	0.582
Correlation between old and new maps	0.981

Table 6–1: Summary of the phase refinement statistics after 12 cycles of solvent flattening with histogram matching and application of Sayres equations of the MIR map. R is defined as $\frac{||F_{obs}|-|F_{calc}||}{|F_{obs}|}$; FOM is the figure of merit.

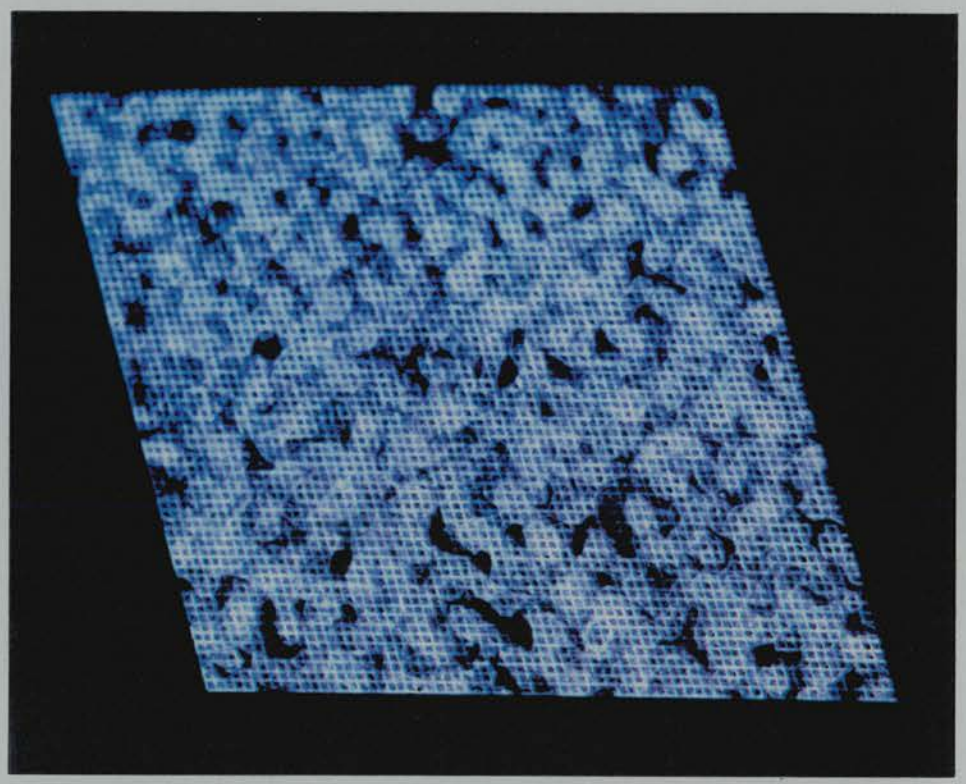
The accuracy of the modified phases is more difficult to assess. The phases after solvent flattening should be defined more accurately than the MIR phases and therefore the figure of merit should increase. Also, the degree of noise in the solvent regions of the map should be reduced and in the case of noncrystallographic symmetry averaging, the correlation between noncrystallographically related subunits should increase. However, all of these changes are caused by the application of the density modification in the first instance. The best measure of the success of density modification is an increase in continuity and a decrease in noise in the electron density map. Statistics for the solvent flattening, histogram matching and application of Sayre's equations to the DlADH MIR map are shown in table 6–1.

The MIR map before solvent flattening was noisy and the protein/solvent boundary was ill-defined (see Figure 6–2). A significant improvement can be seen after 12 cycles of refinement.

6.2.3 Further refinement of heavy atom positions

Density modification can be considered as a step in phase refinement. As such it can be incorporated into a cycle where the heavy atom parameters are refined using phases that have been improved by density modification. This should

a.



b.

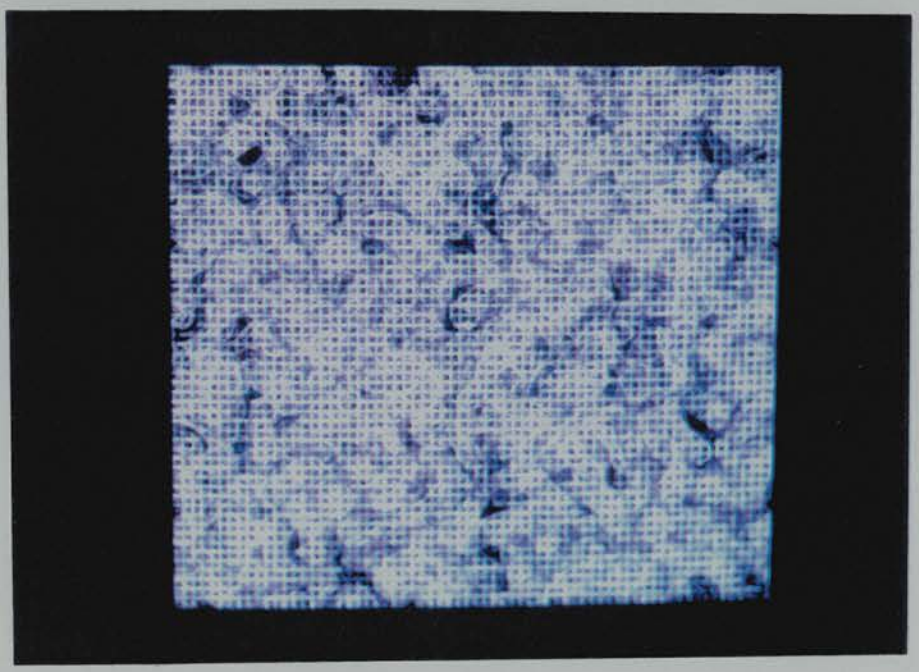
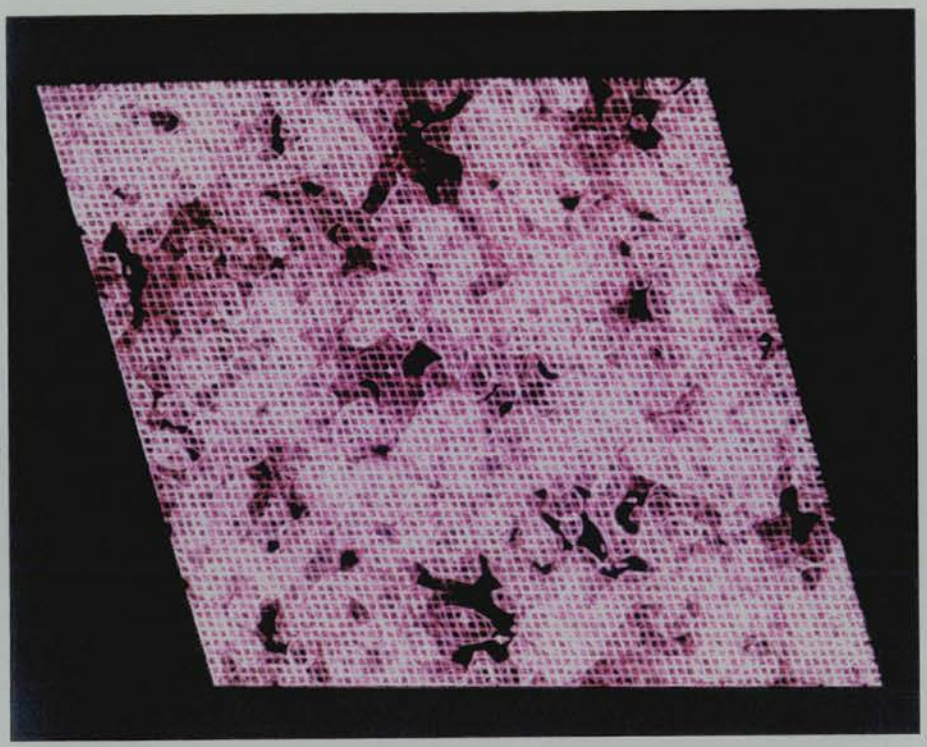


Figure 6–2: Electron density map calculated from the initial MIR phases before density modification a) veiwed down the b-axis b) viewed down the a-axis.

a)



b)

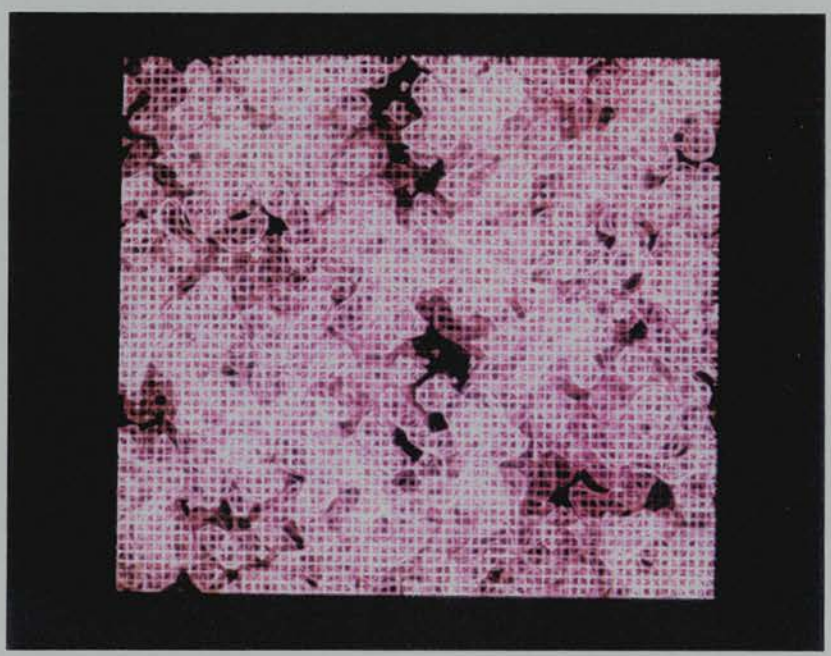


Figure 6–3: The MIR electron density map after 12 cycles of solvent flattening with histogram matching and application of Sayre's equations; a) viewed down the b axis b) viewed down the a-axis.

Data set	Heavy atom position			Occupancy	Anomalous occupancy	B-factor
GHG	0.092	0.378	0.235	1.168	1.145	37.327
	0.349	0.296	0.289	6.424	0.361	35.031
	0.574	0.069	0.387	11.958	0.419	0.225
	0.417	0.366	-0.026	0.198	0.174	-93.413
	0.538	0.120	0.940	0.757	-1.369	2.553
GCMN	0.574	0.069	0.388	26.342	-0.863	-5.464
	0.351	0.293	0.289	14.724	0.237	34.152
	0.400	0.477	0.523	1.124	-2.177	-33.892

Table 6–2: Heavy atom parameters after successive cycles of solvent flattening and refinement against solvent flattened phases. Compare with table 5–6.

	MIR parameters		After SF	
	GCMN	GHG	GCMN	GHG
No. refs	4684	5363	4684	5363
Total phasing power	1.4	1.3	1.4	1.3
Total Rcullis (%)	70	73	69	72
< FOM > (acentric)	0.430		0.441	

Table 6–3: Heavy atom refinement statistics with and without refinement against phases improved by solvent flattening

result in better starting MIR phases being obtained and therefore subsequent cycles of density modification will have a greater chance of producing an interpretable map.

The original MIR map was modified using solvent flattening, histogram matching and Sayre's equations. The original heavy atom parameters were then refined using MLPHARE against the solvent flattened phases. The heavy atom positions did not shift significantly but there were some large changes in the values of the B-factors (see Table 6–2). The process of phase refinement against solvent flattened phases converged after 3 cycles, of solvent flattening (12 cycles) followed by heavy atom refinement (4 cycles). The change in refinement statistics as a result of this refinement are shown in table 6–3. This table should be compared with the statistics obtained in the original MIR refinement (see Table 5–7). The phase statistics as a function of resolution are shown in figures 6–4 to 6–7. The MIR map before and after this phase refinement (see Figures 6–2 and 6–8) are seen to be very similar. There is a slight decrease in the degree of noise in the final map.

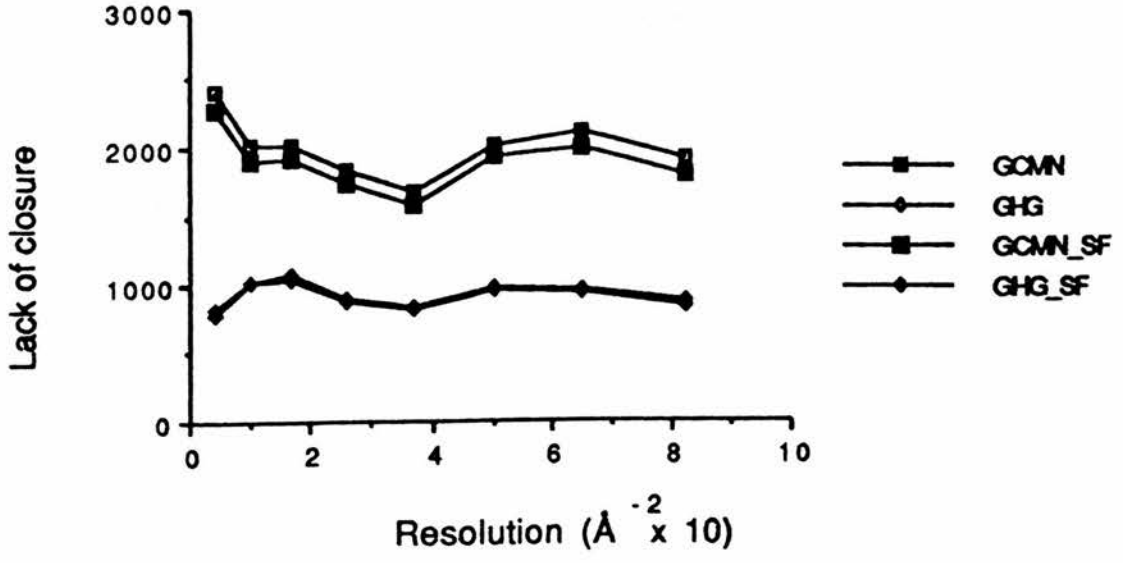


Figure 6–4: Lack of closure (acentric zone) as a function of resolution for GCMN, GHG derivatives and for derivatives after refinement against solvent flattened phases. Where the lack of closure is $|F_{PHobs} - F_{PHcalc}|$.

Although small improvements can be seen between original MIR statistics and MIR statistics after refinement against solvent flattened phases, the degree of improvement in the electron density map is small. Analysis of the refinement statistics as a function of resolution shows that at very low resolution i.e. 15 Å the solvent flattened phase statistics are worse than those from the original MIR map. At higher resolution that statistics are slightly improved.

The MIR map calculated from the heavy atom parameters refined using solvent flattened phases (see Figure 6–8) was subjected to a further 12 cycles of solvent flattening with histogram matching and with Sayre's equations applied. The statistics for this process are shown in table 6–4 and the resulting map is shown in figure 6–9. This map is very different to the map obtained by density modification of the original MIR map (see Figure 6–3). It appears that although successive cycles of density modification and heavy atom refinement have a small effect on the resulting electron density maps, the change in the initial phase set as a result of this refinement is sufficient to make a significant difference to the

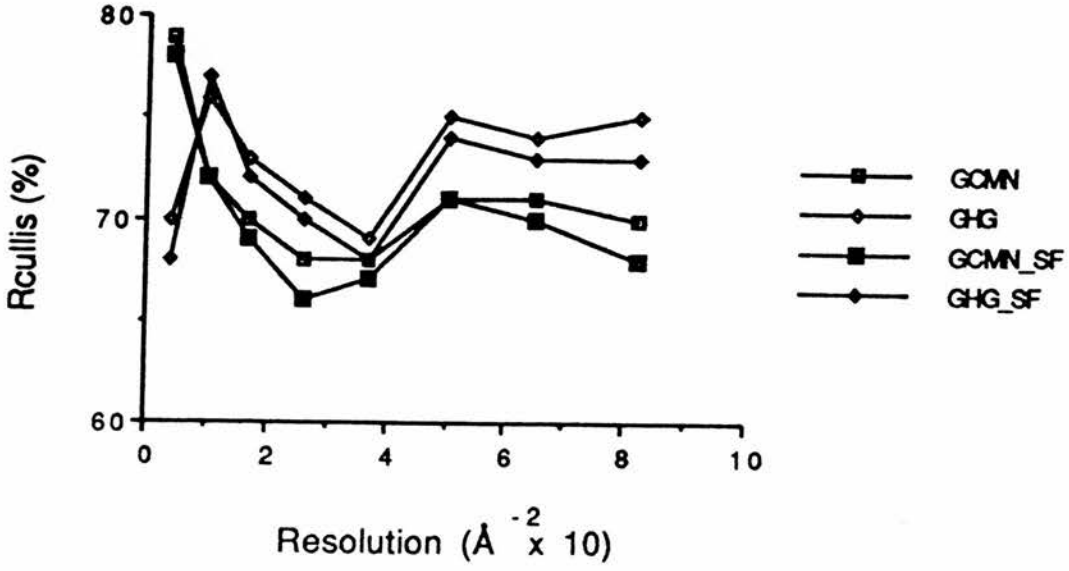


Figure 6–5: Rcullis as a function of resolution for GCMN, GHG derivatives and for the derivatives after refinement against solvent flattened phases. Where Rcullis is defined as $\frac{\Sigma||F_{PH}+/-F_P|-F_{Hcalc}|}{\Sigma|F_{PH}-F_P|}$

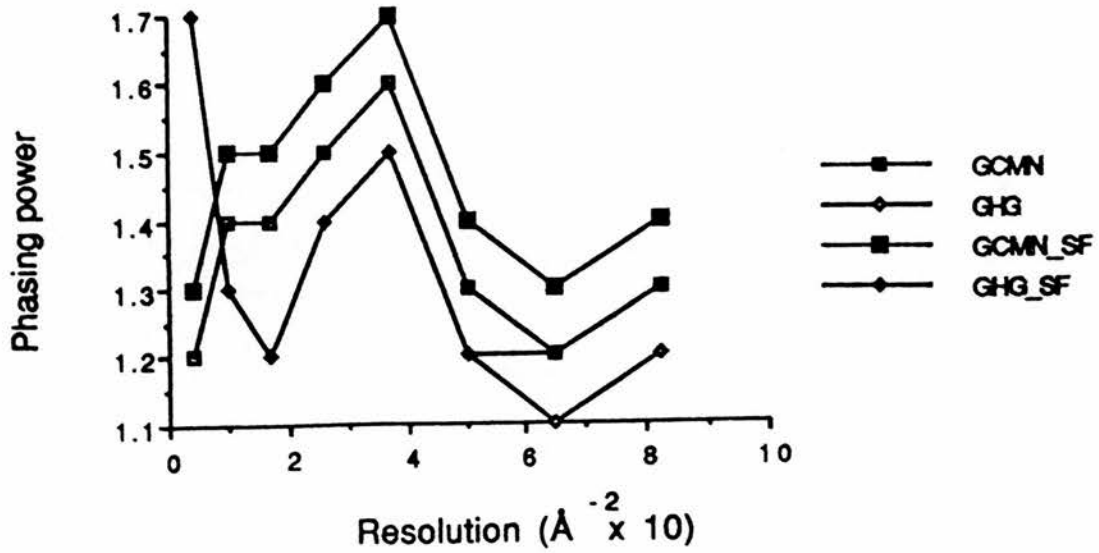


Figure 6–6: Phasing power as a function of resolution for GCMN, GHG derivatives and after refinement against solvent flattened phases. Phasing power F_{Hcalc}/lack of closure

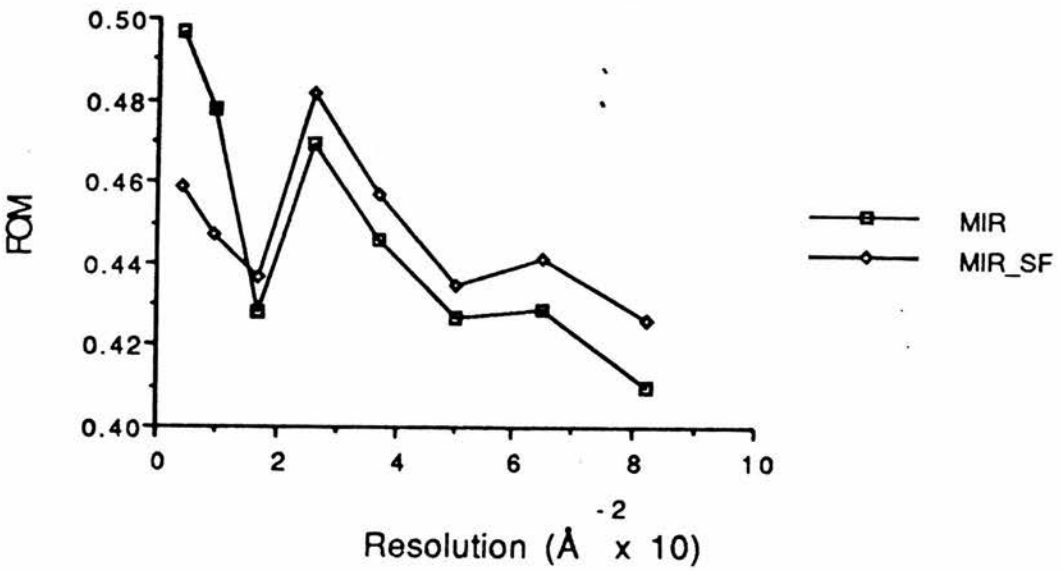
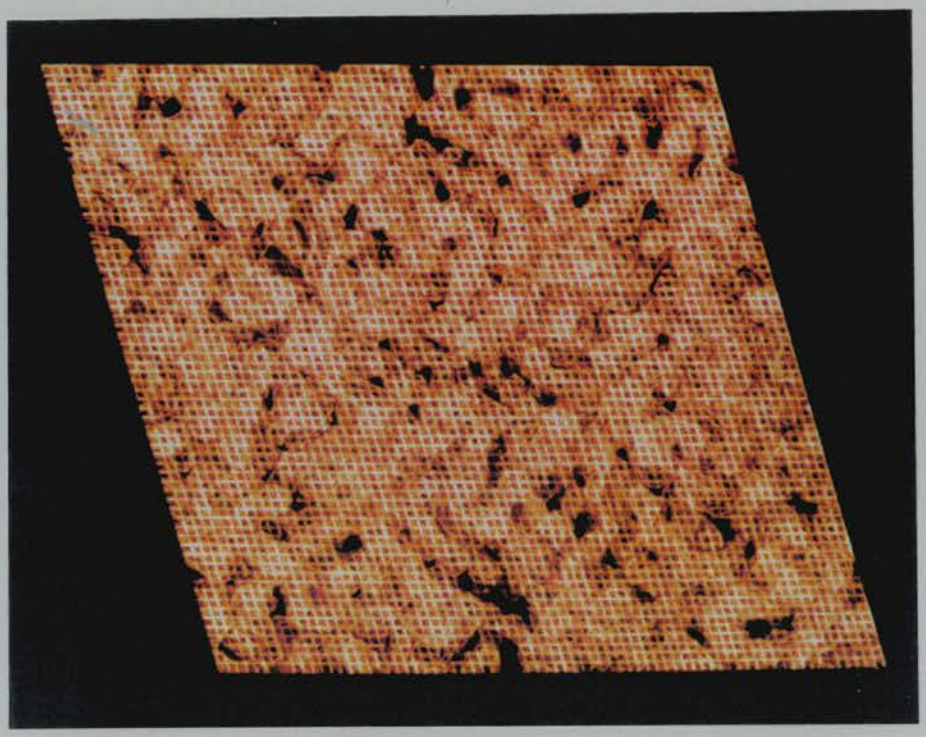


Figure 6–7: Figure of merit (FOM) as a function of resolution and after refinement against solvent flattened phases.

a



b

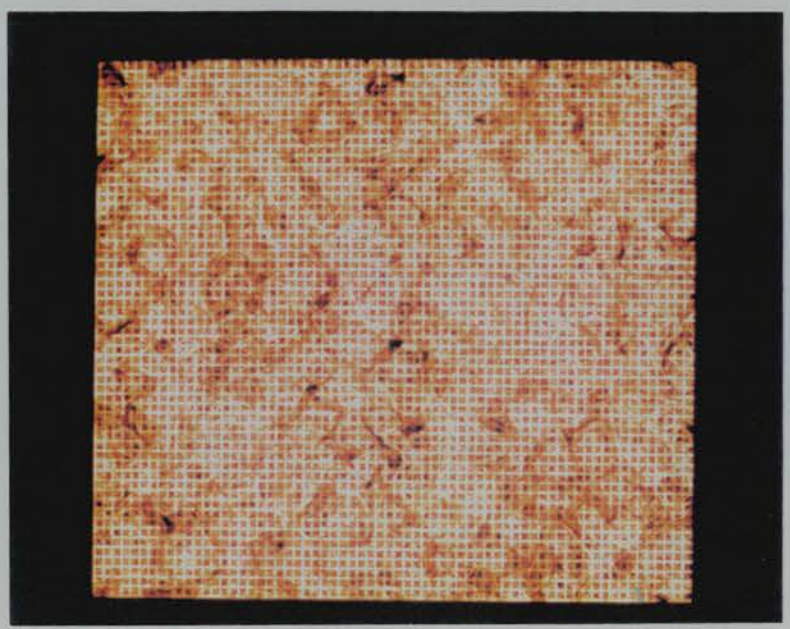


Figure 6–8: MIR map (200 - 3.5 Å) after heavy atom parameters have been refined against density modified phases; a) viewed down b-axis b) viewed down a-axis. Compare to figure 6–2.

No. of cycles of DM	40
No. of cycles of heavy atom refinement	9
No. reflections	6205
R cycle 1	54.4 %
< FOM > (acentric)	0.45
Correlation between F_{obs} and F_{calc}	0.56
Correlation between old and new maps	0.79
R final	67.8 %
< FOM > (acentric)	0.98
Correlation between F_{obs} and F_{calc}	0.84
Correlation between old and new maps	0.94

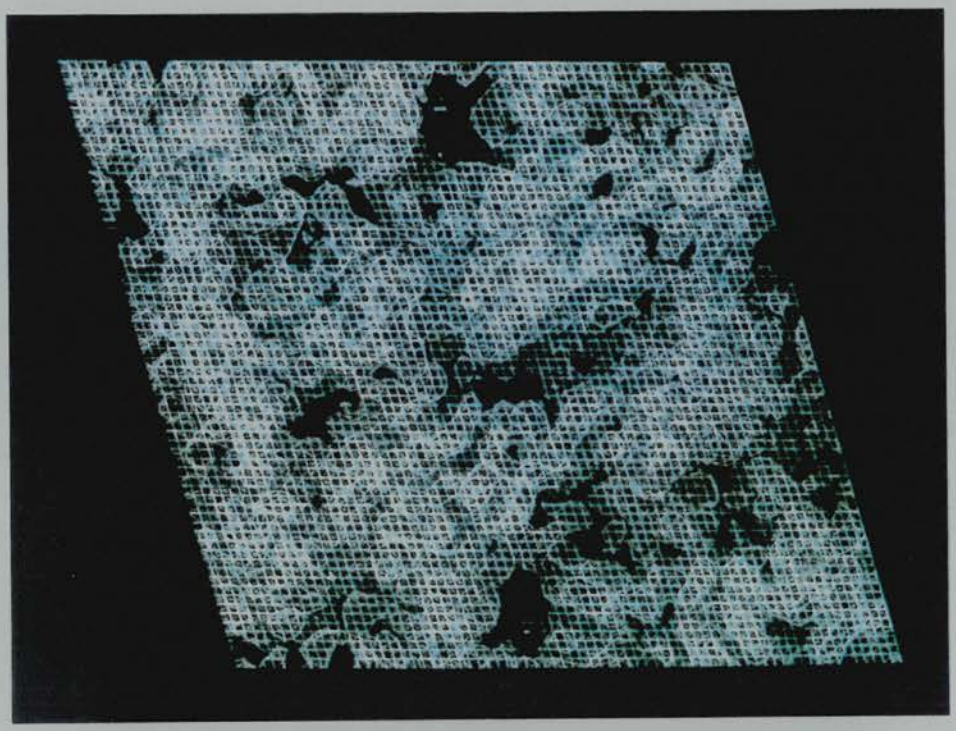
Table 6–4: Summary of the phase refinement statistics after 40 cycles of density modification with refinement of heavy atom parameters against the modified phases every 12 cycles of solvent flattening.

density modification process. The original MIR map and the MIR map after refinement against solvent flattened phases look very similar, but after solvent flattening they have different features. The latter map shows more continuity.

6.2.4 Molecular averaging

Molecular averaging is a powerful method of utilizing a redundancy in the unit cell which results from the presence of noncrystallographic symmetry. The matrix relating the subunits can be determined by looking at the positions of noncrystallographically related heavy atom positions. It did not appear that the heavy atom positions were related by the noncrystallographic symmetry. Since the initial MIR map did not reveal the molecular boundaries of the subunits even at low resolution (100 - 6 Å), solvent flattening, histogram matching and Sayre's equations were applied to the initial MIR map. However, even after 40 cycles of density modification, with heavy atom refinement, the map was not good enough to identify the individual subunits with any certainty. At this stage it was thought unwise to use the matrix determined from the molecular replacement solution to do noncrystallographic averaging, since it might bias the map towards the molecular replacement solution.

a



b

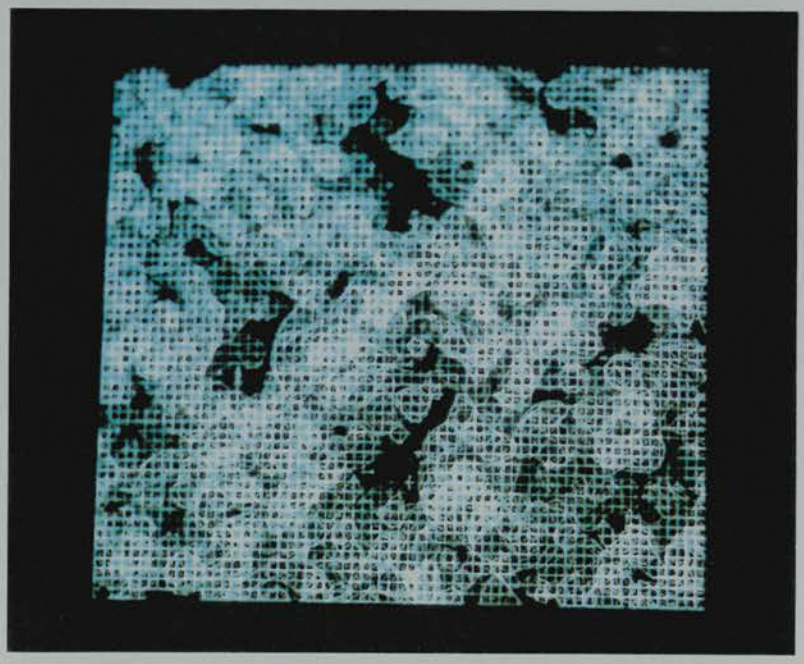


Figure 6–9: Electron density map (200 - 3.5 Å), a) viewed down the b-axis and b) viewed down the a-axis. The map is the result of successive cycles of heavy atom refinement and solvent flattening with histogram matching and application of Sayre's equations

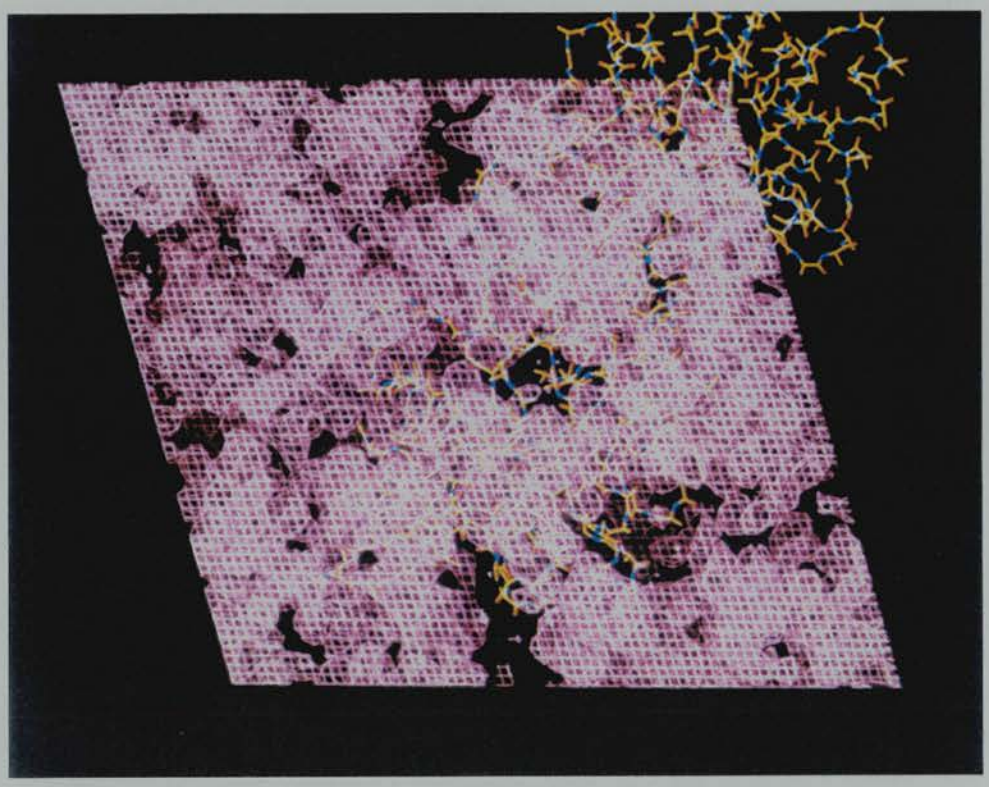


Figure 6–10: Coordinates from MR solution superimposed upon the DM MIR map, viewed down the b-axis.

6.2.5 Comparing the solvent flattened MIR map with the solvent flattened MR solution

The coordinates for the HSD structure, determined from the molecular replacement study were superimposed upon the MIR map that had been subjected to 12 cycles of solvent flattening. Figure 6–10 shows that although there is some overlap most of the HSD structure seems to be displaced relative to the MIR solvent flattened map.

Solvent flattening, histogram matching and Sayre's equations were applied to the weighted MR map. The structure factors were calculated, using SFALL, from the atomic positions determined by the molecular replacement solution. Weights for these structure factors were calculated using SIGMAA. The weighted $3F_o$-$2F_c$

No. of cycles of DM	12
Resolution range of maps	100 - 3.5 Å
No. of reflections	6127
R cycle 1	63.6 %
< FOM >	0.18
Correlation between F_{obs} and F_{calc}	0.56
Correlation between old and new maps	0.79
R final	71.0 %
< FOM >	0.948
Correlation between F_{obs} and F_{calc}	0.882
Correlation between old and new maps	0.981

Table 6–5: Statistics after 12 cycles of density modification of the MR map

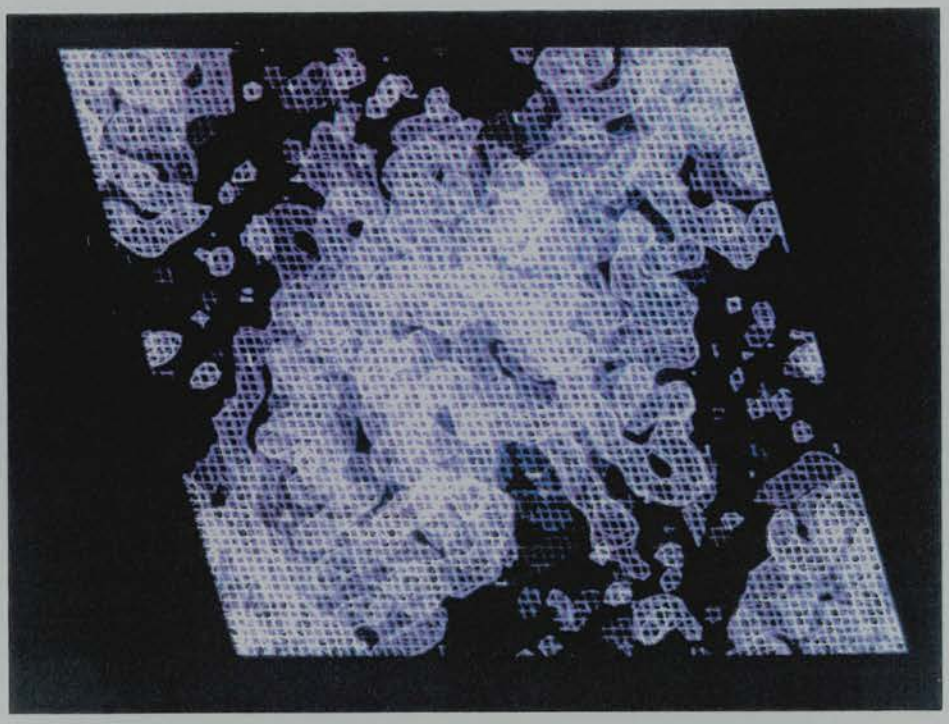
map which was highly biased towards the starting model was then subjected to 12 cycles of solvent flattening with histogram matching and with the application of Sayre's equations. The statistics for the solvent flattening are shown in table 6–5. Sections from the initial map and final map are shown in figure 6–11 and figure 6–12.

Automatic chain tracing using BONES a program within O (Jones *et al.*, 1991) was carried out on the MIR solvent flattened map and the MR solvent flattened map, the overlap of these traces is shown (see Figure 6–13), the continuity in both maps seems to lie in the same region of the cell. The MR map is the more interpretable but there is some concern that this map is biased towards the structure of the input model.

6.3 Discussion

Solvent flattening is a powerful and useful way of improving the initial phases determined by MIR or MR. It is particularly powerful when the solvent content of the crystal is high. Here significant improvements in the electron density map were observed for a moderately low solvent content, only 44%. Solvent flattening of the MIR map resulted in an obvious improvement of the initial MIR map but the resulting map was not good enough to give an unambiguous chain trace. This was due to the inaccuracies of the initial MIR phases. Poor starting phases

a



b

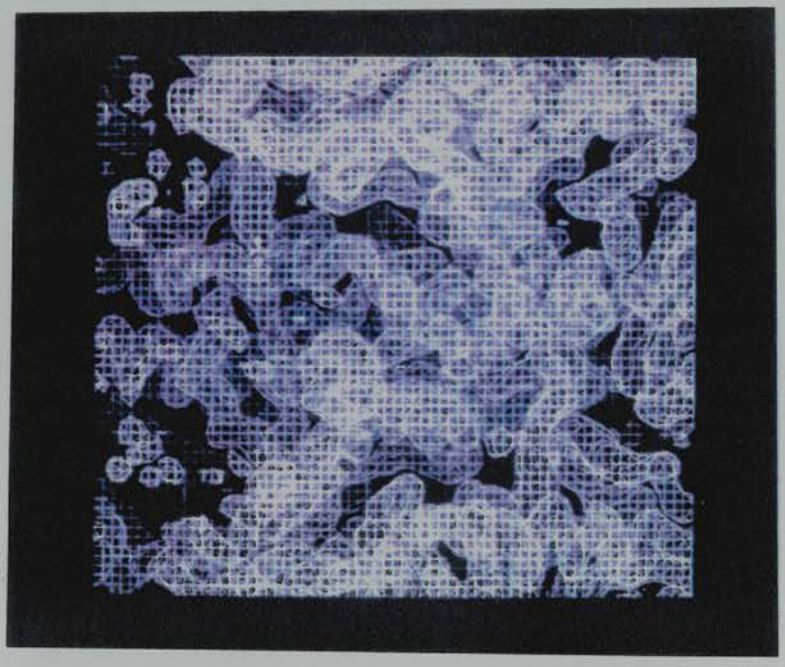
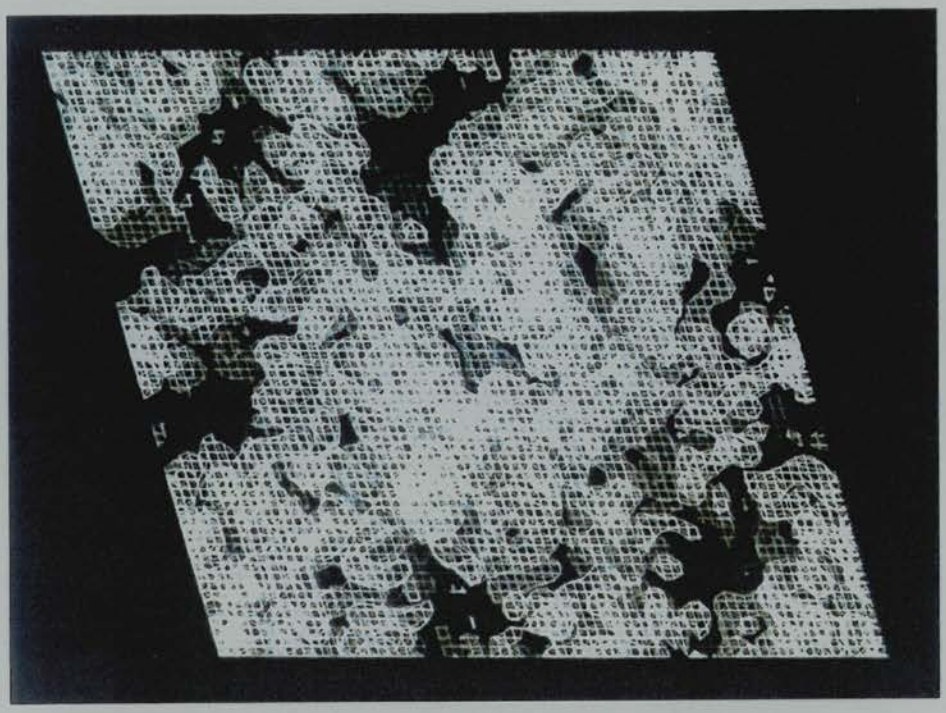


Figure 6–11: Weighted electron density map (100 - 3.5 Å), a) viewed down b-axis b) viewed down a-axis.

a



b

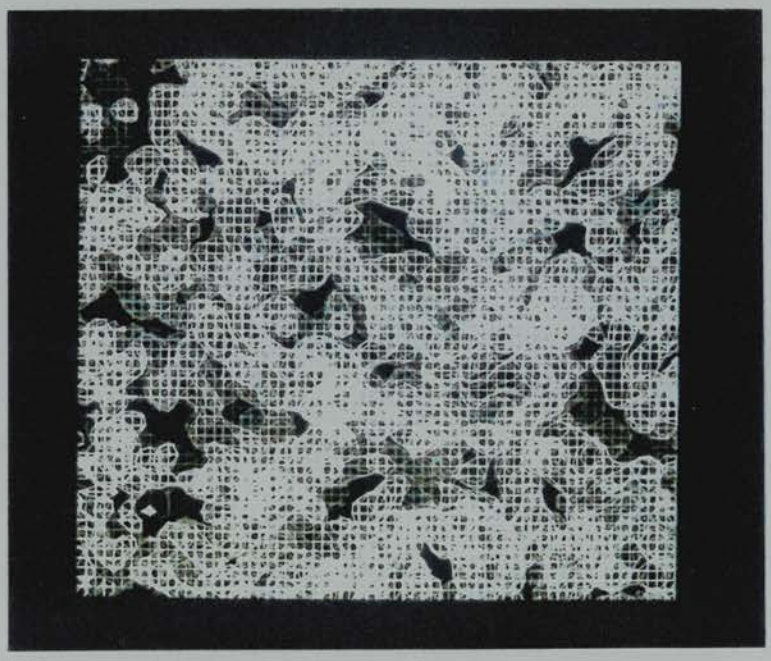


Figure 6–12: MR map (100 - 3.5 Å) after solvent flattening with histogram matching and application of Sayre's equations a) viewed down b-axis and b) viewed down a-axis.

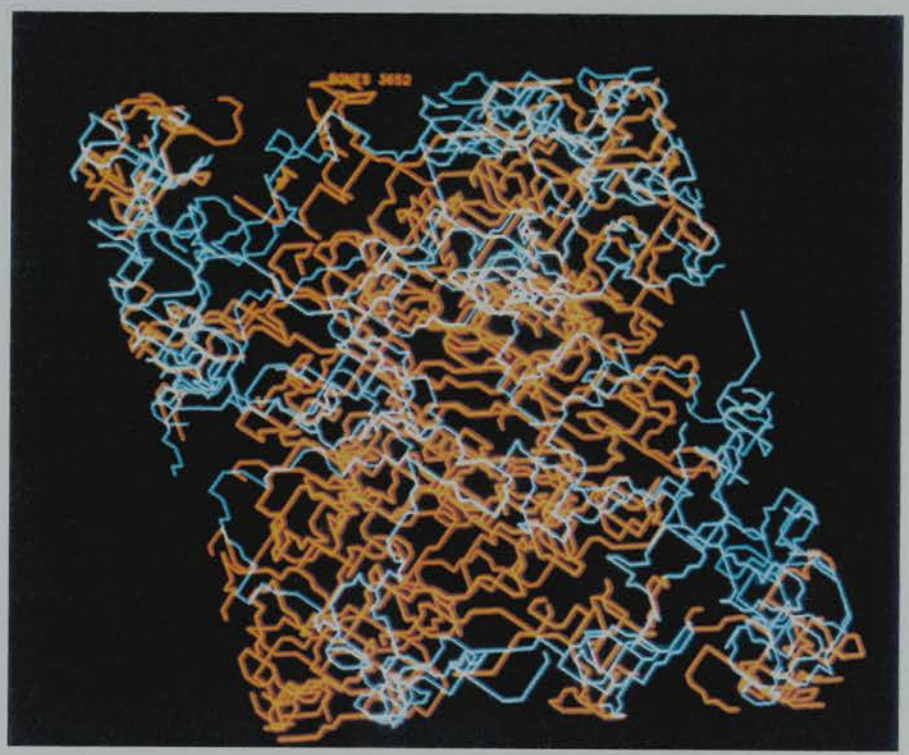


Figure 6–13: Overlap of automatic chain traces on density modified molecular replacement (red) and isomorphous replacement (blue) maps. The traces are viewed down the b axis and the whole unit cell is shown.

result in noisy and inaccurate MIR maps and this can lead to an unsatisfactory determination of the envelope parameters.

These solvent flattened phases have been used in alternate cycles to improve the heavy atom parameters and hence improve the initial MIR phases. Although changes in heavy atom parameters and refinement statistics are subtle the MIR map, density modification of these maps is seen to give an improve electron density map (see Figure 6–9).

Examination of the automatic chain traces carried out on the density modified MIR and MR maps, seem to show that the main regions of continuity in both maps lie in the same part of the unit cell. Therefore, the repeated solvent flattening and heavy atom refinement seem to be converging the MIR map towards the map calculated from the molecular replacement.

6.3.1 Future work

With both MIR and MR maps converging to give similar patterns of electron density, it implies that the MR solution is correct. This can be further tested by using the MR solution to determine a matrix which describes the noncrystallographic symmetry, the matrix can then be used to carry out noncrystallographic symmetry averaging. If the map is reduced to noise and there is no continuous and related regions of density, then this implies that the molecular replacement solution is incorrect or inaccurate.

Interpretation of the above maps should not be undertaken until more phase information is obtained i.e. more heavy atom derivatives are found or further molecular replacement studies have been carried out.

Chapter 7

Discussion and Conclusions

The crystallographic studies carried out on the alcohol dehydrogenase enzyme from *Drosophila* have succeeded in producing a map which shows some promising features. However, at this stage an unambiguous interpretation of the map is not possible. This chapter summarizes the major findings of these crystallographic studies and presents some suggestions for future work.

7.1 Crystallization

Crystallization experiments have determined the necessity of having DTT in all crystallization buffers to obtain high quality protein crystals, suitable for high resolution X-ray diffraction work (form B crystals). In the absence of DTT, form A crystals grow which suggests that there is a change either in the conformation of the protein or in its crystal packing, when one or both of the cysteines in the DlADH enzyme are oxidized. These cysteines are not involved in catalysis (Chen *et al.*, 1990) but recent experiments have indicated that the reduced form of the cysteines is necessary for DmADH stability (Prozorovski *et al.*, 1992). Stability studies have also shown that NAD^+ has a stabilising effect on DmADH (Ribas de Pouplana *et al.*, 1991). The presence of NAD^+ in crystallization buffers promotes the formation of form A crystals and it is proposed that the addition of NAD^+ causes a change similar to that imposed in the absence of DTT which promotes form A crystal formation. This would imply that regions of the protein involved in forming crystal contacts necessary for the formation of form B crystals are near to or form part of the dinucleotide binding site. Also, that the oxidized form of the cysteines in some way obstructs the dinucleotide binding site, either as a result of aggregation of the DADH molecules or by an S-linked low molecular weight ligand (Prozorovski *et al.*, 1992). These observations are consistent with the involvement of Cys-218 (DmADH nomenclature) in cofactor binding, as suggested by Chen *et al.* (1990) and secondary structure predictions which show the two free sulphydryls situated in exposed regions of the structure.

The form B crystals have proved to be suitable for high resolution X-ray

crystallography studies. High resolution data (better then 2 Å) have now been collected on these crystals

Autoindexing of the data collected on the form B crystals indexed a large unit cell, a = 81.24, b = 55.75, c = 109.60 Å and β = 94.26°. However, analysis of the h + l = odd reflections indicated that these reflections were very weak. Hence a smaller cell, a = 70.60, b = 55.75, c = 65.74 Å and β = 106.95° was valid. The small cell is related to the large cell by a rotation about the b-axis. The small cell has been used in all subsequent data analysis.

7.2 Molecular replacement studies

Molecular replacement studies were carried out using a polyalanine model of the HSD, Q-axis dimer. A molecular replacement solution has been determined by constraining the dimer axis of the search model to lie along the noncrystallographic axis found by self-rotation studies. Measurements of the correctness of the molecular replacement study are given by RF values, R-factors, correlation factors and TF values, all of which are affected by the completeness of the search model. Therefore, it is difficult to assess the quality of the solution found using a polyalanine chain. The $3F_o$ - $2F_c$ map calculated from the molecular replacement solution was not good enough to allow an accurate chain trace to be made, although in parts there are regions of continuity, the side chain density. There are two possible reasons why the map is unsatisfactory:

- The phasing power of a polyalanine chain is too poor to give good phases, although polyalanine chains have been used successfully in a number of cases.

- The molecular replacement solution is an artifact and the continuity in the resulting map is due to model bias.

A promising result from the molecular replacement solution was that the phases calculated from the MR solution were used successfully to locate the heavy atom

positions using difference Fourier techniques. However, this is not an infallible test (Dodson, 1992).

Packing diagrams for this solution show a large solvent channel. However the packing diagram is constructed using the HSD search model with loops 202-234 and 237-255 missing; it is possible that when present these loops bridge the channel. The packing diagrams show that the dimers are stacked tightly in the direction of the b-axis with the NAD^+ binding site of each HSD molecule packed against the bottom of the symmetry related molecule. This may explain why the presence of NAD^+ in the crystallization buffers prevent the formation of the form B crystals. Also, why soaking of NAD^+ into pre-existing crystals causes the crystals to crack after 2-3 days.

7.3 Isomorphous replacement

The major problem with using the isomorphous replacement method to determine phases for the DlADH data is that it is a trial and error approach which requires that data collection facilities and X-ray quality crystals are readily available. This was not the case, data collection time and protein supply were scarce. The study was further hindered by the apparent sensitivity of the crystals to soaking with heavy atoms. Initially, the preparation of heavy atom derivatives targeted the free sulphydryl groups. This approach proved to be a problem because of the presence of DTT in the mother liquor. Removal of DTT seemed to disorder the crystals.

Several heavy atom derivative data sets have been collected but the degree of heavy atom substitution is weak. Two reasonable mercury derivatives have been obtained but the major sites in both derivatives are common sites.

7.4 Map improvement

The MIR and MR electron density maps were subjected to density modification using SQUASH. A significant improvement in the map quality was observed. As expected, parts of the map showed an increase in continuity and noise within solvent regions were reduced. A trace of the continuity of the map, using an automatic chain tracing package was carried out on both maps. The continuity in both the solvent flattened MIR and MR maps appeared to lie in the same region suggesting that the MIR and MR map were converging to the same solution. The density within the map was not good enough to identify the two monomers, therefore noncrystallographic symmetry averaging of the MIR map was not possible.

7.5 Future work

Crystallization

Crystallization of DlADH has been optimized but the limiting factor seems to be obtaining a ready supply of protein. Further studies could be undertaken to crystallize alcohol dehydrogenase from *Drosophila melanogaster.* This is a sensible approach since most studies have been carried out on *D. melanogaster* but may not be feasible since crystallizing *D. melanogaster* ADH has proved problematic (G.K. Chambers, personal communication).

Molecular replacement

- Current molecular replacement studies have gone as far as possible with the model available. It may soon be possible to use the complete highly refined model of HSD as a search model.

- An alternative approach would be to use the Rossmann fold of a medium chain dehydrogenase as a search model. However, this would rely on

determining the medium chain dehydrogenase which was most closely related to the short chain dehydrogenase family and to the *Drosophila* alcohol dehydrogenase family in particular.

- As more structures for short chain dehydrogenases become available further molecular replacement studies become possible.

Isomorphous replacement

- Check that removal of DTT from the crystals does not result in nonisomorphism. Test a range of soaking times to find optimum time for back soaking.

- To find more isomorphous derivatives. A promising complex seems to be platinum chloride but more effort is needed to optimize soaking conditions. Another possibility is to target the methionine residue using methylmercury nitrate.

Map improvement

Map improvement works but the results depend on the reliability of the initial phases and these were improved by successive cycles of solvent flattening and heavy atom parameter refinement. Future work should be directed towards improving the initial phases and the best method of doing this is to find more heavy atom derivatives. Once a reasonable map has been determined, where the two monomers in the asymmetric unit can be identified, noncrystallographic symmetry averaging can be used to improve the map.

Chapter 8

Bibliography

Adams, M. J. (1987). Oxido-reductases - Pyridine Nucleotide-dependent enzymes. *Enzyme Mechanisms.* Ed. M.I. Page and A. Williams. 477-506.

Albalat, R. and Gonzàlez-Duarte, R. (1990) Nucleotide sequence of the Adh gene of *Drosophila lebanonesis. Nucleic Acids Res.* **18**, 6706.

Albalat, R., Gonzàlez-Duarte, R. and S. Atrian (1992). Protein engineering of *Drosophila* alcohol dehydrogenase. The hydroxyl group of Tyr^{152} is involved in the active site of the enzyme. *FEBS lett.* **308**, 235-239.

Alberola, J., Sanchez, A. and Fontdelvila, A. (1987). Adh expression in species of the *mulleri* subgroup of *Drosophila. Biochem. Genet.* **25**, 729-738.

Arndt, U.W. and Gilmore, D.J. (1979). X-ray television area detectors for macromolecular structural studies with synchrotron radiation sources. *J. Appl. Cryst.* **12**, 1-9.

Arndt, U.W. and Wonacott, A.J. (1977). *The rotation method in crystallography.* North-Holland Publishing Co.

Atkinson, P.W., Mills, L.E., Starmer, W.T. and Sullivan, D.T. (1988). Structure and evolution of the Adh genes of *Drosophila mojavensis. Genetics.* **120**, 713-723.

Atrian, S. and Gonzalez-Duarte, R. (1985). Purification and characterization of alcohol dehydrogenase from *Drosophila hydei*: conservation in the biochemical features of the enzyme of several species of *Drosophila. Biochem. Genet.* **23**, 891-911.

Baker, P.J., Britton, K.L., Rice, D.W., Rob, A. and Stillman, T.J. (1992) Structural consequences of sequence patterns in the finger print region of the nucleotide binding fold. Implications for nucleotide specificity. *J. Mol. Biol.* **228**, 662-671.

Batterham, P., Gritz, E., Starmer, W.T. and Sullivan, D.T. (1983). Biochemical characterization of the products of the Adh loci of *Drosophila mojavensis. Biochem. Genet.* **21**, 871-883.

Benyajati, C., Place, A.R., Powers, D.A. and Sofer, W. (1981). Alcohol dehydrogenase gene of *D. melanogaster*: Relationship of intervening sequences to functional domains in protein. *Proc. Natl. Acad. Sci. U.S.A.* **78**, 2717-2721.

Beurkens, P.T., Gould, R.O., Bruins Slot, H.J. and Bosman, W.P. (1987). Translation functions for the positioning of a well orientated molecular fragment. *Z. Krist.* **179**, 127-159.

Bijvoet, J.M. (1949). Phase determination in direct Fourier synthesis of crystal structures *Proc. Neth. Acad. Sci.* **52**, 313-314.

Blake, C.F. (1983). Exons - present from the beginning ? *Nature* **306**, 535-537.

Blow, D.M. (1959) The structure of haemoglobin VII. Determination of phase angles in the non-centrosymmetric [100] zone. *Proc. Roy. Soc.* **A247**, 302-307.

Blow, D.M. (1985). Introduction to rotation and translation functions. *Proceedings from the Daresbury study weekend on Molecular Replacement.* 2-7.

Blow, D.M. and Crick, F.H.C. (1959). The treatment of errors in the isomorphous replacement method. *Acta. Cryst.* **12**, 794-802.

Blundell, T.L. and Johnson, L.N. (1976). *Protein crystallography.* Academic Press.

Bodmer, M. and Ashburner, M. (1984). Conservation and change in the DNA sequences coding for alcohol dehydrogenase in sibling species of *Drosophila*. *Nature.* **309**, 425-430.

Borras, T., Persson, B. and Jörnvall, H. (1989). Eye lens δ crystallin relationships to the family of 'long chain' alcohol/polyol dehydrogenases protein trimming and conservation of stable parts. *Biochem.* **28**, 6133-6139.

Brändèn, C.-I. (1980). Founding fathers and families. *Nature.* **345**, 607-608.

Bricogne, G. (1982). Multiple isomorphous replacement: the problem of parameter refinement from acentric reflections. *Computational crystallography.* Ed. Sayre 223-230.

Bricogne, G. (1991). A maximum-likelihood theory of heavy-atom parameter refinement in the isomorphous replacement method. *Proceedings of the CCP4 study weekend on isomorphous replacement and anomalous scattering.* 60-68.

Brändén, C.-I. (1990). Founding fathers and families. *Nature.* **346**, 607-608.

Brändén, C.-I. and Tooze, J. (1991). *Introduction to protein structure.* Garland Publishing, Inc. New York and London.

Brünger, A. (1990). Extension of Molecular replacement: A new search Strategy based on Patterson correlation refinement. *Acta. Cryst.* **46**, 46-57.

Buehner, M. and Hecht, H.J. (1985) Real space vs. reciprocal. *Proceedings from the Daresbury study weekend on Molecular Replacement.* 62-69.

Carrea, G., Pasta, P and Vacchio, G. (1984). Effect of the lyotropic series of anions on denaturation and renaturation of 20β-hydroxysteroid dehydrogenase. *Biochim. Biophys. Acta.* **784**, 16-23.

Carter, C.W. (Jr.) (Ed.)(1990). Protein and nucleic acid crystallization. *Methods: A companion to methods in enzymology* **1(1)**.

Carter, C.W. (Jr.) and Carter, C.W. (1979). Protein crystallization using incomplete factorial experiments. *J. of Biol. Chem.* **254(23)**, 12219-12223.

Chambers, G.K. (1988). The *Drosophila* alcohol dehydrogenase gene-enzyme system. *Advances in Genetics.* **25**, 39-107.

Chambers, G.K. (1991). Mini-review: Gene expression, adaption and evolution in higher organisms, evidence from studies of *Drosophila* alcohol dehydrogenases. *Comp. Biochem. Physiol.* **99B(4)**, 723-730.

Chambers, G.K., Laver, W.G., Campbell, S. and Gibson, J.B. (1981a). Structural analysis of an electrophoretically cryptic alcohol dehydrogenase variant from an Australian population of *Drosophila melanogaster. Proc. Natl. Acad. Sci.* **34**, 625-637.

Chambers, G.K., Wilks, A.V. and Gibson, J.B. (1981b). An electrophoretically cryptic alcohol dehydrogenase variant in *Drosophila melanogaster* III. Biochemical properties and comparison with common enzyme forms. *Aust. J. Biol. Sci.* **34**, 625-637.

Chambers, G.K., Wilks, A.V. and Gibson, J.B. (1984). Variation in the biochemical properties of the *Drosophila* alcohol dehydrogenase allozymes. *Biochem. Genet.* **22**, 153-168.

Chen, Z., Lu., L., Shirley, M., Lee, W.R. and Chang, S.H. (1990). Site-directed mutagenesis of Gly-14 and two 'critical' Cys residues in *Drosophila* alcohol dehydrogenase. *Biochem.* **29**, 1112-1118.

Chen, Z., Lee, W.R. and Chang, S.H. (1991). Role of aspartic acid 38 in the cofactor specificity of *Drosophila* alcohol dehydrogenase. *Eur. J. Biochem.* **202**, 263-267.

Chen, Z., Lin, Z.G., Lee, W.R. and Chang, S.H. (1992). Tyrosine-152 of *Drosophila* alcohol dehydrogenase is an essential residue for activity. *The FASEB Journal.* **6(1)**, Abstract No. 2643.

Chothia C. and Lesk A. (1986). The relation between the divergence of sequence and structure in proteins. *The EMBO Journal.* **5(4)**, 823-826.

Cohn, V.H., Thompson, M.A. and Moore, G.P. (1984) Nucleotide sequence comparison of the Adh gene in three drosophilids. *J. Mol. Evol.* **20**, 31-37.

Cowtan, K. (1991). Sayre's equation and histogram methods for phase refinement and extension of protein structures. *Crystallographic computing 5, from chemistry to biology.* Papers presented at the international school on crystallographic computing held at Bischenberg, France. Edited by D. Moras, A.D. Podjarny and J.C. Thierry. International Union of Crystallography Oxford University Press. 373-381.

Coyne, J.A. and Kreitman, M. (1986) Evolutionary genetics of two sibling species, *Drosophila simulans* and *Drosophila sechellia. Evolution.* **40**, 673-691.

Crowther, R.A. (1972). Fast rotation function. *The molecular replacement method: a collection of papers on the use of non-crystallographic symmetry.* Ed. M.G. Rossmann. Publishers Gordon and Breach. New York. 173-178.

Crowther, R.A. and Blow, D.M. (1967). Method for positioning a known molecule in an unknown crystal structure. *Acta. Cryst.* **23**, 544-548.

Cura, V., Podjarney, A.D., Khrishnaswarmy, Rees, B., Rondeau, J.M., Tete, F., Mourey, L., Samama, J.P. and Moras, D. (1991). Heavy atom refinement against solvent-flattened and local symmetry averaged phases. *Proceedings of the CCP4 study weekend on isomorphous replacement and anomalous scattering*, 107-115.

Dalziel, K. (1963). Kinetic studies of liver alcohol dehydrogenase and pH effects with coenzymes preparations of high purity. *J. Biol. Chem.* **238**, 2850-2858.

Danielsson, O. and Jörnvall, H. (1992). "Enzymogenesis": Classical liver alcohol dehydrogenase origin from the glutathione-dependent formaldehyde dehydrogenase line. *Proc. Natl. Acad. Sci. USA.* **89**, 9247-9251.

David, J.R., v. Herrewege, J., Monclus, M. and Prevosti, A. (1979). High ethanol tolerance in two distantly related *Drosophila* species: A probable case of recent convergent adaptation. *Comp. Biochem. Physiol.* **63C**, 53-56.

Day, T.H., Hillier, P.C. and Clark, P. (1974). Properties of genetically polymorphic isoenzymes of alcohol dehydrogenase in *Drosophila melanogaster. Biochem. Genet.* **11**, 141.

van Deldon, W. (1982). The alcohol dehydrogenase polymorphism in *Drosophila melanogaster*: selection at an enzyme locus. *Evol. Biol.* **15**, 187-222.

Diamond, R. (1969). Profile analysis in single crystal diffractometry *Acta. Cryst.* **A25**, 43-55.

Dodson, E.J. (1976). A comparison of different heavy atom refinement procedures. *Computing Methods in Crystallography.* Ed. Ahmed. 259-268.

Dodson, E.J. (1985) Molecular replacement: the methods and its problems. *Proceedings of the Daresbury Study weekend on Molecular Replacement.* 33-45.

Dodson, E.J. (1992). From the molecular replacement solution to the refined structure. *Proceedings of the CCP4 study weekend on Molecular replacement.* 84-86.

Ducruix, A. and Giegé, R. (1991). Crystallization of nucleic acids and proteins. A practical approach. IRL Press.

Dunn, G.R., Wilson, T.G. and Jacobsen, K.B. (1969). Age-dependent changes in alcohol dehydrogenase in *Drosophila. J. exp. Zool.* **171**, 185-189.

Durbin, R.M., Burns, R., Moulai, J., Metcalf, P., Freymann, D., Blum, M., Anderson, J.E., Harrison, S.C. and Wiley, D.C. (1986). Protein, DNA and virus crystallography with a focused imaging proportional counter. *Science.* **232**, 1127-1132.

Eagles, P.A.M., Johnson, L.N. and Joynson, M.A. (1969) Subunit structure of aldolase: chemical and crystallographic evidence. *J. Mol. Biol.* **45**, 533-544,

Eklund, H., Nordström, B., Zepperzauer, E., Söderlund, G., Ohlsson, I., Biowe, T. and Brändèn, C.-I. (1974). The structure of horse liver alcohol dehydrogenase. *FEBS Lett.* **44(2)**, 200-204.

Eklund, H., Nordström, B., Zepperzauer, E., Söderlund, G., Ohlsson, I., Biowe, T., Söderberg, B.O., Tapia, O., Brändèn, C.-I. and Akeson, A. (1976). Three dimensional structure of horse liver alcohol dehydrogenase at 2.4Å resolution. *J. Mol. Biol.*, **102**, 27-59.

Eklund, H. and Brändén, C.-I. (1987). Alcohol Dehydrogenases. *Biological Macromolecules and Assemblies (Active site of enzymes)* **3**. Ed. Jurnak, F.A. and M^{c}Pherson, A. 74-142.

Eklund, H., Müller-Wille, P., Horjales, E., Futer, O., Holmguist, B., Vallee, B.L., Hoog, J.-O., Kaiser, R. and Jörnvall, H. (1990). Comparisons of three classes of human liver alcohol dehydrogenases. Emphasis on different binding pockets. *Eur. J. Biochem.* **193**, 303-310.

Ensor, C.M. and Tai, H.-H. (1991). Site-directed mutagenesis of the conserved tyrosine 151 of human placental NAD^+-dependent 15-hydroxysteroid dehydrogenase yields a catalytically inactive enzyme. *Biochem. Biophys. Res. Comm.* **176(2)**, 840-845.

Evans, P.R., Farrants, G.W., Laurence, M.C. and Shirakihara, Y. (1985). Low resolution structures of two forms of phosphofructokinase. *Proceedings from the Daresbury study weekend on Molecular Replacement*, 53-55.

Evans, P.R. (1991) Refinement of heavy-atom parameters and isomorphous phasing. *Proceedings of the Daresbury Study weekend on Isomorphous Replacement.* 49-59.

Fersht, A. (1977). *Enzyme structure function and mechanism.* Freeman. New York.

Fisher, J.A. and Maniatis, T. (1985). Structure and transcription of the *Drosophila mulleri* alcohol dehydrogenase gene. *Nucleic Acids Res.* **13**, 6899-6977.

Fitzgerald, P.M.D. (1988). MERLOT, An integrated package of computer programs for the determination of crystal structures by molecular replacement. *J. Appl. Cryst.* **21**, 273-278.

Ford, G.C. (1974). Intensity determinations by profile fitting applied to precession photographs. *J. Appl. Cryst.* **7**, 555-564.

Fox, G.C. and Holmes, K.C. An alternative method of solving the layer scaling equations of Hamilton, Rollet and Sparks. *Acta. Cryst.* **20**, 886-891.

Fujinaga, M. and Read, R.J. (1987). Experiences with a new Translation-function Program. *J. Appl. Cryst.* **20**, 517-521.

Geer, B.W., Heinstra, P.W.H., Kapoun, A.M. and Van der Zel, A. (1990). Alcohol dehydrogenase and alcohol tolerance in *Drosophila melanogaster. Ecological and Evolutionary genetics of Drosophila* Ed. Barker, J.S.F. *et al.* Chapter 13, 231-252. Plenum press, New York.

Gernstein, M. and Chothia C. (1991). Analysis of protein loop closure. Two types of hinges produce one motion in lactate dehydrogenase. *J. Mol. Biol.* **220**, 133-149.

Ghosh, D., Weeks, C.M., Grochulski, P., Duax, D.L., Erman, M., Rimsay, R.L. and Orr, J.C. (1991). Three-dimensional structure of holo 3α,20β-hydroxysteroid dehydrogenase: A member of a short chain dehydrogenase family. *Proc. Natl. Acad. Sci. U.S.A.* **88**, 10064-10068.

Gibson, J.B., Chambers, G.K., Wilks, A.V. and Oakeshott, J.B. (1980). An electrophoretic cryptic alcohol dehydrogenase variant of *Drosophila melanogaster*. 1. Activity ratios, thermostability, genetic locatization and comparison with two other thermostable variants. *Aust. J. Biol. Sci.* **33**, 479-489.

Gordon, E.J., Bury, S.M., Sawyer, L., Atrian, S. and Gonzalez-Duarte, R. (1992). Preliminary X-ray crystallographic studies on alcohol dehydrogenase from *Drosophila. J. Mol. Biol.* **227**, 356-358.

Green, D.W, Ingram, V.M., and Perutz, M.F. (1954). The structure of hemoglobin IV sign determination by the isomorphous replacement method. *Proc. R. Soc. London. Ser.* **A225**, 287-307.

Gunnarsson, P.O., Pettersson, G., and Zepperzauer, E. (1974). Inhibition of horse liver alcohol dehydrogenase by $Pt(CN)_4^{2-}$ and $Au(CN)_2^-$. *Eur. J. Biochem.* **43**, 479.

Harada, Y., Lifchitz, A., Berthou, J. and Jolles, P. (1981). A translation function combining packing and diffraction information: An application to lysozyme (high-temperture form). *Acta. Cryst.* **A37** 398-406.

Heinstra, P.W.H., Eisses, K.Th., Choonen, W., Aben, W.J.A., De Winter, A.J., Van der Hurst, D.J., Van Marrewijk, W.J.A., Beenakkers, A.M.Th., Scharloo, W. and Thörig, G.E.W. (1983). A dual function of alcohol dehydrogenase in *Drosophila. Genetica.* **60**, 129-137.

Heinstra, P.W.H.,Scharloo, W. and Thorig, G.E.W. (1988) Alcohol dehydrogenase polymorphism in *Drosophila.* Enzyme kinetics of product inhibition. *J. Molec. Evol.* **28**, 145-150.

Hendrickson, W.A., Horton, J.R. and LeMaster, D.M. (1990) Selenomethionyl proteins produced for analysis by multiwavelength anomalous diffraction (MAD): a vehicle for direct determination of three-dimensional structure. *The EMBO Journal.* **9(5)**, 1665-1672.

Hendrickson, W.A and Lattmann, E.E. (1970). Representation of phase probability distributions for simplified combinations of independent phase information. *Acta. Cryst.* **B26**, 136-143.

Hendrickson, W.A. and Ward, K.B. (1976). A packing function for delimiting the allowable locations of crystallized macromolecules. *Acta. Cryst.* **A32**, 778.

Hovik, R., Winberg, J.-O. and McKinley-McKee, J.S. (1984). *Drosophila melanogaster* alcohol dehydrogenase, substrate stereospecificity of the Adh^F alleloenzyme. *Insect. Biochem.* **14(3)**, 345-351.

Howard, A.J., Guilliland, G.L., Finzel, B.C., Poulos, T.L., Ohlendorf, D.H. and Salemme, F.R. (1987). The use of an imaging proportional counter in macromolecular crystallography. *J. Appl. Cryst.* **20**, 383-387.

Howell, P.L. and Smith, G.D. (1992). Identification of heavy atom derivatives by normal probablility methods *J. Appl. Cryst.* **25**, 81-86.

Huber, R. (1985). Experiences with the application of Patterson search techniques. *Proceedings of the Daresbury study weekend on Molecular Replacement.* 58-61.

Hurley, T.D., Bosron, W.F., Hamilton, J.A., Amzel, L.M. (1991). Structure of human $\beta_1\beta_1$ alcohol dehydrogenase: catalytic effects of non-active-site substitutions *Proc. Natl. Acad. Sci.* **88**, 8149-8153.

Inoue, T., Sunagawa, M., Mori, A., Imai, C., Fukuda, M., Takagi, M and Yano, K. (1989). Cloning and sequencing of the gene encoding th 72-kilodalton dehydrogenase subunit of alcohol dehydrogenase from *Acetobacter aceti. J. Bacteriol.* **171**, 3115-3122.

Irie, S., Doi, S., Yorifuji, T., Takagi, M. and Yano, K. (1987). Nucleotide sequenceing and characterisation of the genes encoding benzene oxidation enzymes of *Pseudomonas putida. J. Bacteriol.* **169**, 5174-5179.

Jancarik, J. and Kim, S.-H. (1991). Sparse matrix sampling: a screening method for crystallization of proteins *J. Appl. Cryst.* **24**, 409-411.

Jany, K.-D., Ulmer, W., Fröschle, M. and Pfleiderer, G. (1984). Complete amino acid sequence of glucose dehydrogenase from *Bacillus megaterium. FEBS Lett.* **165**, 6-10.

Jones, T.A., Zou, J.-Y., Cowan, S.W. and Kjeldgaard, M. (1991). Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Cryst.* **A47**, 110-119.

Jones, E.Y. and Stuart, D.I. (1992). Methods of structural analysis of proteins. Part 1 - protein crystallography. *Protein Engineering. A practical appraoch series.* Eds. A.R. Rees, M.J.E. Sternberg and R. Wetzel. IRL Press. 3-32.

Jörnvall, H. (1977). Differences between alcohol dehdyrogenases. Structural properties and evolutionary aspects. *Eur. J. Biochem.* **72**, 443-452.

Jörnvall, H., von Bahr-Lindström, H., Jany, K-D., Ulmer, W. and Fröschle, M. (1984a). Extended superfamily of short-alcohol-polyol-sugar dehydrogenases: structural similarities between glucose and ribitol dehydrogenases. *FEBS Lett.* **165(2)**, 190-195.

Jörnvall, H., von Bahr-Lindström, H. and Jeffery, J. (1984b). Extensive variations and basic features in the alcohol dehydrogenase - sorbitol dehydrogenase family. *Eur. J. Biochem.* **140**, 17-23.

Jörnvall, H., Persson, M. and Jeffery, J. (1981). Alcohol and polyol dehydrogenases are both divided into two protein types, and structural properties cross-relate the different enzyme activities within each type. *Proc. Natl. Acad. Sci. U.S.A.* **78(7)**, 4226-4230.

Jörnvall, H., Persson, B., Krook, M. and Kaiser, R. (1990). Alcohol dehydrogenases *Biochem. Soc. Trans.* **18**, 169-171.

Juan, E. and González-Duarte, R. (1980). Purification an enzyme stability of alcohol dehydrogenases from *Drosophila simulans*, *Drosophila virilis* and *Drosophila melanogaster* adh^S. *Biochem. J.* **189**, 105-110.

Juan, E. and González-Duarte, R. (1981). Determination of some biochemical and structural features of alcohol dehydrogenase from *Drosophila simulans* and *Drosophila virilis*: comparison of their properties with the *Drosophila melanogaster* Adh^S enzyme. *Biochem. J.* **195**, 61-69.

Kabsch, W. (1988a). Automatic indexing of rotation diffraction patterns. *J. Appl. Cryst.* **21**, 67-71.

Kabsch, W. (1988b). Evaluation of single-crystal X-ray diffraction data from a position-sensitive detector. *J. Appl. Cryst.* **21**, 916-924.

Katz, B.A., Ollis, D. and Wyckoff, H.W. (1985). Low resolution crystal structure of muconolactone isomerase *J. Mol. Biol.* **184**, 311-318.

Knowles, B.B and Friström, J.W. (1967). Electrophoretic behaviour of 10 enzyme systems in larval integument of *Drosophila melanogaster. J. Insect. Physiol.* **13**, 731-737.

Krook, M., Marekov, L. and Jörnvall, H. (1990). Purification and structural characterisation of placental NAD^+-linked 15-hydroxyprostaglandin dehydrogenase. The primary structure reveals the enzyme to belong to the short-chain alcohol dehydrogenase family. *Biochem.* **29**, 738-743.

Krook, M., Prozorovski, V., Atrian, S., Gonzàlez-Duarte and Jórnvall, H. (1992). Short-chain dehydrogenases proteolysis and chemical modification of prokaryotic $3\alpha,20\beta$-hydroxysteroid, insect alcohol and human 15-hydroxyprostaglandin dehydrogenases. *Eur. J. Biochem.* **209**, 233-239.

Kvassman, J. and Pettersson, G. (1980). Unified mechanism for proton-transfer reactions affecting the catalytic activity of liver alcohol dehydrogenase. *Eur. J. Biochem.* **103**, 565-575.

Kvassman, J. and Pettersson, G. (1980). Effect of pH on the binding of decanoate and trifluoroethanol to liver alcohol dehydrogenase. *Eur. J. Biochem.* **103**, 557-564.

Langs, D.A. (1985). Translation functions: The minimization of structure-independent spurious maxima. *Acta. Cryst.* **A41**, 578-582.

Lattman, E.E. (1985). Use of the rotation and translation functions. *Methods in Enzymology* **115**, 55-77.

Leslie, A.G.W. (1988). A reciprocal space algorithm for calculating molecular envelope using the algorithm of B.C. Wang. *Proceedings of the Daresbury study weekend on Improving protein phases.* 25-31.

Leslie, A.G.W. (1991). Molecular data processing. *Crystallographic computing 5, from chemistry to biology.* Papers presented at the international school on crystallographic computing held at Bishenberg, France. Edited by D. Moras, A.D. Podjarny and J.C. Thierry. International union of crystallography Oxford university press. 50-61.

Lipson, H. and Cochran, W. (1966). *The determination of crystal structures.* 3rd edition, Bell and Sons, London.

Lui, J., Duncan, K. and Walsh, C.T. (1989). Nucleotide sequence of a cluster of *Echerichia coli* enterobactin biosynthesis genes: Indentification of *EntA* and purification of its product 2,3-dihydro-2,3-dehydroxybenzoate dehydrogenase. *J. Bacteriol.* **171**, 791-798.

Marekov, L., Krook, M. and Jörnvall, H. (1990). Prokaryotic 20β-hydroxysteroid dehydrogenase is an enzyme of the 'short chain, non-metalloenzyme' alcohol dehydrogenase type. *FEBS Lett.* **266**, 51-54.

Magdoff, B.S., Crick, F.H.C. and Luzzati, V. (1956). The three-dimensional Patterson function of ribonuclease II. *Acta. Cryst.* **9**, 186-162.

Matthews, B.W. (1968). Solvent content of protein crystals. *J. Mol. Biol.* **33**, 491-497.

McKinley-McKee, J.S., Winberg, J.-O. and Pettersson, G. (1991). Mechanisim of action of *Drosophila melanogaster* alcohol dehydrogenase. *Biochem. Int.* **25(5)**, 879-885.

McPherson, A. (1982). Preparation and analysis of protein crystals. Wiley, New York.

McPherson, A. (1985). Use of polyethylene glycol in the crystallization of macromolecules *Methods in enzymology* **114**, 120-125. Wyckoff, H.W., Hirs, C.H.W., and Timasheff, S. Academic press, Orlando.

McPherson, A. (1985). Crystallization of proteins by variation of pH or temperature. *Methods in enzymology* **114**, 125-128.

McPherson, A. (1990). Current approaches to macromolecular crystallization. *Eur. J. Biochem.* **189**, 1-23.

Messerschmidt, A. and Pflugrath, J.W. (1987). Crystal orientation and X-ray pattern predication routines for area-detector diffractometer systems in macromolecular crystallography. *J.Appl.Cryst.* **20**, 306-315.

Moxon, L.N., Holms, L.S. Parsons, P.A., Irving, M.G. and Doddrell, D.M. (1985). Purification and molecular properties of alcohol dehydrogenase from *Drosophila melanogaster*: evidence from NMR and kinetics studies for function as an aldehyde dehydrogenase *Comp. Biochem. Physiol.* **B 80**, 525-535.

Nagai, K., Evans, P.R., Li, J., Oubridge, Ch. (1991). Phase determination using mercury derivatives of the engineered cysteine mutants. *Proceedings of the CCP4 study weekend: Isomorphous replacement and anomalous scattering.* 141-149.

Nyborg, P. and Wonacott, A.J. (1977). The rotation method in crystallography. Editors U.W. Arndt and A.J. Wonacott. North Holland Publishing Co.

Oakeshott, J.G., Chambers, G.K., East, P.D., Gibson, J.B. and Barker, J.S.F. (1982). Evidence for a genetic duplication involving alcohol dehydrogenase gene is *Drosophila buzzattii* and related species. *Aust. J. Biol.* **29**, 365-373.

Otwinowski, Z. (1991). Maximium likelihood refinement of heavy atom parameters. *Proceedings of the CCP4 study weekend on isomorphous replacement.* 80-86.

Patterson, A.L. (1934). A Fourier series method for the determination of the components on interatomic distances in crystals. *Phys. Rev.* **46**, 372.

Persson, B., Krook, M. and Jörnvall, H. (1991). Characterisation of short chain alcohol dehydrogenases and related enzymes. *Eur. J. Biochem.* **200**, 537-543.

Petsko, G.A. (1985). Preparation of isomorphous heavy-atom derivatives *Methods in Enzymology.* **114**, 147-155.

Petsko, G.A., Phillips, D.C., Williams, R.J.P. and Wilson, I.A. (1978). On the protein crystal chemistry of chloroplatinate ions: general principles and interactions with triose phosphate isomerase. *J. Mol. Biol.* **120**, 345-359.

Pettersson, G. (1986) Zinc Enzymes (Bertini, I., Luchinat, C., Maret, W. and Zeppezauer, M. Eds.)

Phillips, D.C. (1966) *Advances of structural research by diffraction methods* Interscience. Ed. R. Brill and R. Mason. New York and London.

Place, A.R., Powers, D.A. and Sofer, W. (1980) *Drosophila melanogaster* alcohol dehydrogenase does not require metals for catalysis. *Fed. Proc.* **39**, 1640.

Podjarny, A.D. and Rees, B. (1991). Density modification: theory and practice. *Crystallographic computing 5, from chemistry to biology.* Papers presented at the international school on crystallographic computing held at Bishenberg, France. Edited by D. Moras, A.D. Podjarny and J.C. Thierry. International union of crystallography Oxford university press. 361-372.

Prozorovski, V., Krook, M., Atrian, S., Gonzàlez-Duarte, R. and Jörnvall, H. (1992). Identification of reactive tyrosine residues in cysteine - reactive dehydrogenases. *FEBS lett.* **304(1)**, 46-50.

Rao, S.N., Jih, J.-H. and Hartsuck, J.A. (1980). Rotation function space groups. *Acta. Cryst.* **A36**, 878-884.

Retzio, A. and Thatcher, D.R. (1981). Characterization of the Adh^{F1} and Adh^{US} alleloenzymes of *Drosophila melanogaster* (fruitfly) alcohol dehydrogenase. *Biochem. Soc. Trans.* **9**, 298-299.

Ribas de Pouplana, Ll., Atrian, S., González-Duarte, R., Fothergill-Gilmore, L.A., Kelly, S.M. and Price, N.C. Structural properties of Long- and Short-Chain Alcohol dehydrogenases: Contribution of NAD^+ to stability. *Biochem. J.* (1991), **276**, 433-438.

Ribas de Pouplana, Ll. and Fothergill-Gilmore, L.A. (1993) Two distinct classes of short chain dehydrogenases suggested by site-directed mutagenesis and computer modelling. In preparation.

Richardson, J.S. (1984). The anatomy and taxonomy of protein structure. *Adv. Prot. Chem.* **34**, 167-339.

Ringe, D., Petsko, G.A., Yamakura, F., Suzuk, K. and Ohmori, D. (1983) Structure of iron superoxide dismutase from *Pseaudomonas ovalis* at 2.9Å resoltution. *Proc. Natl. Acad. Sci.* **80**, 3879.

Rossmann, M.G. (1974). Chemical and biological evolution of a nucleotide-binding protein. *Nature* **250**, 194-199.

Rossmann, M.G. (1979). Processing oscillation diffraction data for very large unit cells with an automatic convolution technique and profile fitting. *J. Appl. Cryst.* **12**, 225-238.

Rossmann, M.G. and Blow, D. (1961). The refinement of structures partially determined by the isomorphous replacement method. *Acta. Cryst.* **14**, 641-647.

Rossmann, M.G. and Blow, D. (1962). The detection of sub-units within the crystallographic asymmetric unit. *Acta. Cryst.* **15**, 24.

Rossmann, M.G., Adams, M.J., Buehner, M., Ford, G.C., Hackert, M.L., Liljas, A., Rao, S.T., Banaszak, L.J., Hill, E., Tsernoglou, D. and Webb, L. (1973). Structural constraints of possible mechanisms of lactate dehydrogenase as shown by high resolution studies of the apoenzyme and a variety of enzyme complexes. *J. Mol. Biol.* **76**, 533-537.

Rossmann, M.G., Lilias, A., Brändén, C.-I. and Banaszak, L.J. (1975). Evolutionary and structural relationships among dehydrogenases. *The Enzymes.* **3** 3rd Edition. Ed. Boyer P.D. (Academic, New York), 61-102.

Rowan R.G and Dickerson, W.J. (1988). Nucleotide sequence of the genomic region encoding alcohol dehydrogenase in *Drosophila affinidisjuncta. J. Mol. Evol.* **28**, 43-54.

SERC Daresbury laboratory. (1991) Chapter 5: Protein crystallography. *The SRC.*

Schaeffer S.W. and Aquadro, C.F. (1987). Nucleotide sequence of the Adh gene region of the *Drosophila pseudoobscura*: evolutionary change and evidence of an ancient gene duplication. *Genetics.* **117**, 61-73.

Schierbeek, B. (1991) New developements on the Enraf-Nonius detector. *Crystallographic computing 5, from chemistry to biology.* Papers presented at the international school on crystallographic computing held at Bishenberg, France. Edited by D. Moras, A.D. Podjarny and J.C. Thierry. International union of crystallography Oxford university press. 62-68.

Schwartz, M.F. and Jörnvall, H. (1976). Structural analysis of mutant and wild type alcohol dehydrogenases from *Drosophila melanogaster*. ***Eur. J. Biochem.*** **68**, 159-168.

Scopes, R.K. (1983). An iron-activated alcohol dehydrogenase ***FEBS lett.*** **156(2)**, 303-306.

Scrutton, N.S., Berry, A. and Perham, R.N. (1990). Redesign of the coenzyme specificity of a dehydrogenase by protein engineering. ***Nature.*** **343, 38-43.**

Sheldrick, G. (1991). Heavy Atom location using SHELX-90. ***Proceedings from the CCP4 study weekend on isomorphous replacement and anomalous scattering*** 23-38.

Skarzyuski, T., Moody, P.C.E. and Wonacott, A.J. (1987). Structure of holo-glyceraldehyde-3-phosphate dehydrogenase from ***Bacillus stearothermophilus*** at 1.8 Åresolution. *J. Mol. Biol.* **193**, 171-187.

Smith, J.L. (1991). Determination of three-dimensional structure by multiwavelength anomalous diffraction. *Curr. Opinion. Struc. Biol.* **1(6),** 1002-1011.

Sofer, W. and Martin, P. (1987). Analysis of alcohol dehydrogenase gene expression system in *Drosophila. Ann. Rev. Genet.* **21,** 203-225.

Sygusch, J. (1977). Minimum-varience Fourier coefficients from the isomorphous replacement method by least-squares analysis. *Acta. Cryst.* **A33,** 512-518.

Taylor, W.R. (1988). Review: Pattern matching methods in protein sequence comparison and structure prediction. *Protein engineering.* **2(2),** 77-86.

Terwilliger, T.C. and Eisenberg, D. (1983). Unbiased three-dimensional refinement of heavy atom parameters by correlation of origin-removed Patterson functions. *Acta. Cryst.* **A39,** 813-817.

Thatcher, D. (1977). Enzyme stability and proteolysis during the purification of an alcohol dehydrogenase from *Drosophila melanogaster. Biochem. J.* **163,** 317-323.

Thatcher, D.R. and Retzio, A. (1981). Mutations affecting the structure of the alcohol dehydrogenase from *Drosophila melanogaster. Proteins and related subjects* **28**, 157-160.

Thatcher, D.R. and Sawyer, L. (1980). The complete amino acid sequence of three alcohol dehydrogenase alleloenzymes (Adh^{N-11},Adh^{s} and Adh^{UF}) from the fruitfly *Drosophila melanogaster.J. Mol. Biol.* **187**, 875-886.

Thatcher, D.R. and Sheik, R. (1981) The relative comformational stability of the alcohol dehydrogenase alleloenzymes of the fruitfly *Drosophila melanogaster. Biochem. J.* **197**, 111-117.

Theorell, H. and McKinley-McKee, J.S. (1961). Liver alcohol dehydrogenase *Acta. Chem. Scand.* **15**, 1787-1810.

Thörig, G.E.W., Schoone, A.A. and Scharloo, W. (1975). Variations between electrophoretically identical alleles at the alcohol dehydrogenase locus in *Drosophila melanogaster. Biochem. Genet.* **13**, 721-731.

Ursprung, H., Sofer, W. and Burroughs, N. (1970) Ontogeny of tissue distribution of alcohol dehydrogenase in *Drosophila melanogaster. Wilhelm Roux' Arch.* **164**, 201-208.

Vigue, C.L. and Johnson, F.M. (1973) Isoenzyme variability in species of the genus *Drosophila.* VI. Frequency-property-environment relationships of the allelic alcohol dehydrogenase in *Drosophila melanogaster. Biochem. Genet.* **9, 213.**

Vilagelui, L.C. and Gonzalez-Duarte, R. (1984). Alcohol dehydrogenase from *Drosophila funebris* and *Drosophila immigrans*: molecular and evolutionary aspects. *Biochem. Genet.***22**, 797-815.

Villarroya, A., Juan, E., Egestad, B. and Jörnvall, H. (1989). The primary structure of alcohol dehydrogenase from *Drosophila lebanonensis. Eur. Biochem.***180**, 191-197.

Villieux, F.M.D., Groendijk, H., Huitema, F., Swarte, M.B.A., Drenth, J. and Hol, W.G.J. (1988). The use of solvent-flattening procedures in the crystal structure determination of quinoprotein methylamine dehydrogenase. *Proceedings of the CCP4 study weekend on improving protein phases.* 88-100.

Wang, B.C. (1985). Resolution of phase ambiguity in macromolecular crystallography. *Methods in Enzymology* **115**, 90-112.

Weaver, J.R., Andrews, J.M. and Sullivan, D.T. (1989) Nucleotide sequences of the Adh-1 gene of *Drosophila navojoa. Nucleic Acids Res.* **17**, 7524.

Weber, P.C. (1991). Physical principles of protein crystallization *Advances in protein chemistry.* **41**, 1-36.

Weis, W.I. and Brünger, A.T. (1989). Crystallographic refinement by simulated annealing. *The proceedings of the Daresbury study weekend on Molecular simulation and protein crystallography.*

Wierenga, R.K., Terpstra, P. and Hol, W.G.J. (1985). Prediction of the occurence of the ADP-binding $\beta\alpha\beta$-fold in proteins, using an amino acid finger print. *J. Mol. Biol.* **187**, 101-107.

Williamson, V.M. and Paquin, C.E. (1987). Homology of *Saccharomyces cerevisiae ADH4* to an iron-activated alcohol dehydrogenase from *Zymomonas mobilis. Mol. Gen. Genet.* **209**, 374-381.

Winberg, J.-O., Hovik, R., McKinley-McKee, J.S., Juan, E. and Gonzalez-Duarte, R. (1986). Biochemical properties of alcohol dehydrogenase from *Drosophila lebanonensis. Biochem. J.* **235**, 481-490.

Winberg, J.-O. and McKinley-McKee, J.S. (1988a). *Drosophila melanogaster* alcohol dehydrogenase. Biochemical properties of the NAD^+-plus-acetone-induced isoenzymes conversion. *Biochem. J.* **251**, 223-227.

Winberg, J.-O and McKinley-McKee, J.S. (1988b). The ADH^S alleloenzyme of alcohol dehydrogenase from *Drosophila melanogaster*: variation of kinetic parameters with pH. *Biochem. J.* **255**, 589-599.

Winberg, J.-O. and McKinley-McKee, J.S. (1992). **Kinetic interpretations of** active site topologies and residue exchanges in ***Drosophila*** **alcohol dehydrogenase.** *Int. J. Biochem.* **24(2)**, 169-181.

Winberg, J.-O., Thatcher, D.R. and McKinley-McKee, J.S. **(1982a). Alcohol** dehydrogenase from the fruitfly ***Drosophila melanogaster*** **substrate specificity of** the alleloenzyme Adh^S and Adh^{UF}. ***Biochim. Biophys. Acta.*** **704, 7-16.**

Winberg, J.-O., Thatcher, D.R and McKinley-McKee, J.S. **(1982b). Alcohol** dehydrogenase from the fruitfly ***Drosophila melanogaster*** **Inhibition studies of the** alleloenzymes Adh^S and Adh^{UF}. ***Biochim. Biophys. Acta.*** **704, 17-25.**

Winberg, J.-O., Thatcher, D.R. and McKinley-McKee, J.S. **(1983).** ***Drosophila*** *melanogaster* **alcohol dehydrogenase an electrophoretic study of the** Adh^S, Adh^F and Adh^{UF} alleloenzymes. ***Biochem. Genet.*** **21, 63-80.**

Winberg, J.-O., Hovik, R. and McKinley-McKee, J.S. **(1985). The alcohol** dehydrogenase alleloenzymes Adh^S and Adh^F **from the fruitfly** ***Drosophila*** *melanogaster*: **An enzymatic rate assay to determine the active site** concentration. ***Biochem. Genet.*** **23**, 205-216.

Yamada, M. and Saier, M.H. Jr (1987). **Glucitol-specific enzymes of the** phosphotransferase system in ***Escherchi coli.*** **Nucleotide sequence of the gut** operon. *J. Biol. Chem.* **262**, 5455-5463.

Zhang, K.Y.J. and Main, P. (1988). **Histogram matching as a density** modification technique for phase refinement **and extension of protein molecules.** ***Proceedings of the Daresbury study weekend on Improving protein phases.*** **57-64.**

Appendix A

Published paper

Reprinted from *J. Mol. Biol.* (1992) **227**, 356–358

Preliminary X-ray Crystallographic Studies on Alcohol Dehydrogenase from *Drosophila*

Elspeth J. Gordon, Stella M. Bury, Lindsay Sawyer Silvia Atrian and Roser Gonzalez-Duarte

J. Mol. Biol. (1992) **227**, 356–358

Preliminary X-ray Crystallographic Studies on Alcohol Dehydrogenase from *Drosophila*

Elspeth J. Gordon, Stella M. Bury, Lindsay Sawyer†

Department of Biochemistry, University of Edinburgh
Hugh Robson Building, George Square, Edinburgh EH8 9XD, Scotland

Silvia Atrian and Roser Gonzalez-Duarte

Department de Genetica, Universitat de Barcelona
Diagonal 645, 08071 Barcelona, Spain

(Received 11 May 1992; accepted 26 May 1992)

The alcohol dehydrogenase (ADHase) enzyme catalyses the oxidation of alcohols to aldehydes or ketones using NAD^+ as a cofactor. Functional ADHase from *Drosophila lebanonensis* is a dimer, with a monomeric molecular weight of 27,000 and with 254 residues in each polypeptide chain. Crystals of the protein have been grown with and without NAD^+. Two crystal forms have been observed. Most crystals are plate-like, 0·05 mm in their shortest dimension and up to 0·4 mm in their longest dimension. These crystals are generally too small to diffract efficiently using conventional X-ray sources, so preliminary studies were carried out using the Synchrotron Radiation Source at the SERC Daresbury Laboratory. Twinning was a severe problem with this crystal form. The second form is grown in the absence of NAD^+ but with DL-dithiothreitol present. These crystals grow more evenly and diffract to better than 2 Å resolution. They are monoclinic, with cell dimensions, $a = 81{\cdot}24(6)$ Å, $b = 55{\cdot}75(4)$ Å, $c = 109{\cdot}60(7)$ Å and $\beta = 94{\cdot}26(9)°$, space group $P2_1$. There are two dimers in the asymmetric unit, but at low resolution a rotated cell with one dimer per asymmetric unit can be obtained.

Keywords: alcohol dehydrogenase; *Drosophila*; "short-chain" dehydrogenase

Alcohol dehydrogenases, EC 1.1.1.1 (ADHase‡), constitute a relatively complex group of enzymes that catalyse the conversion of primary and secondary alcohols to aldehydes and ketones. The sequences have been divided into four families (Persson *et al.*, 1991): "long chain", "medium chain", "short chain" and an iron-dependent enzyme. The original long chain ADHase family has now been termed the medium chain family because of the discovery of a longer ADHase. The new long chain group is not well characterized (Inoue *et al.*, 1989).

The medium chain group is diverse, members containing typically 350 to 400 amino acids per polypeptide chain. Dehydrogenases of this type have been isolated from mammals, higher plants and yeast. All of these enzymes use NAD^+ or $NADP^+$, frequently but not always contain one or two zinc atoms per subunit and in the active form, are either dimeric or tetrameric. There is little sequence identity within this group. Each monomer is composed of two domains: one binds substrate and the other coenzyme. The coenzyme binding region is located towards the C terminus of the protein and constitutes a Rossmann fold involving approximately 150 residues. Horse liver alcohol dehydrogenase (LADHase) is the only medium chain ADHase for which there is a crystal structure (for a review, see Eklund & Brändén, 1987). Its structure has helped in the postulation of a catalytic mechanism that involves the transfer of a hydride from the substrate to the C-4 atom of the nicotinamide ring of the NAD^+. The family also includes other dehydrogenases, such as lactate dehydrogenase, soluble malate dehydrogenase, glyceraldehyde-3-phosphate dehydrogenase and 6-phosphogluconate dehydrogenase. Crystal structures of these enzymes all reveal the presence of $\beta\alpha\beta$ supersecondary structure in the coenzyme binding

† Author to whom all correspondence should be addressed.

‡ Abbreviations used: ADHase, alcohol dehydrogenase; DTT, dithiothreitol

0022–2836/92/170356–03 $08.00/0

domains while the structures of the catalytic regions vary. An eye lens crystallin has recently been added to this family because of the high sequence similarity and the similarity of its conformational properties to those of the other medium chain enzymes (Borras *et al.*, 1989). This protein, however, does not appear to have any zinc atom bound and no dehydrogenase activity has been shown.

The short chain dehydrogenase family is not as well characterized as the medium chain one. Initially it contained only *Drosophila* ADHase (DADHase) and bacterial ribitol dehydrogenase (Jörnvall *et al.*, 1981), but recently the family has expanded rapidly and it now includes more than 20 enzymes. Active forms of the short chain enzymes, like the medium chain enzymes, are dimers or tetramers. Each monomer has a M_r of around 27,000 with approximately 250 residues in the polypeptide chain, but there are no zinc atoms bound to the protein. NAD^+, but as far as is known not $NADP^+$, is used as a cofactor, although a recent mutagenesis experiment has produced an enzyme capable of using both cofactors (Chen *et al.*, 1991). Only six of the 250 residues are strictly conserved with glycine being the most conserved residue (Persson *et al.*, 1991), the sequence identity between pairs of enzymes is approximately 25%. One crystal structure of a short chain dehydrogenase has been solved recently: that of 3α,20β-hydroxysteroid dehydrogenase has been determined to 2·6 Å (1 Å = 0·1 nm) resolution (Ghosh *et al.*, 1991).

The fourth class of dehydrogenase has been defined from the structure of a prokaryotic iron-dependent enzyme (Scopes, 1983; Jörnvall *et al.*, 1990), but little is known about this class.

Alcohol dehydrogenase from *Drosophila* (DADHase) has been intensively studied for many years with regard to its evolution, genetics and structural features (e.g. see van Deldon, 1982; Chambers, 1988). The primary structures for several *Drosophila* ADHases have been published and sequence comparisons show identities of up to 87%. Secondary structural predictions for the *D. melanogaster* ADHase (Thatcher & Sawyer, 1980) indicate that the monomers are organized into two functional units: a catalytic domain and a cofactor binding domain; the latter being a Rossmann fold, but, in this family, situated in the N-terminal region of the protein.

The present study has been carried out with DADHase from *D. lebanonensis*, a species of the Scaptodrosophila radiation, because it is more stable in solution than the *D. melanogaster* enzyme, which precipitates readily with total loss of enzymic activity. The primary structure of the *D. lebanonensis* enzyme has been determined (Villaroya *et al.*, 1989). The substrate specificities of the known DADHases are the same (Winberg *et al.*, 1986), with a preference for secondary alcohols, contrasting with the zinc-containing medium chain ADHase, which are more active towards primary alcohols.

The protein was purified using a modification of the method of Juan & Gonzalez-Duarte (1980) as described by Ribas de Pouplana *et al.* (1990). Protein (0·9 mg ml^{-1}) in nitrogen-purged storage buffer, 20 mM Tris·HCl (pH 8·6) with 1% isopropanol, 10^{-4} M-DTT and 0·2% mercaptoethanol, was dialysed against 20 mM-Tris·HCl (pH 8·6) containing 10^{-4} M-DTT. The protein solution was concentrated using an Amicon concentrator with YM10 filter, to a final concentration of approximately 5 mg ml^{-1}. Crystals with and without cofactor were prepared using the hanging-drop technique at 10°C, from 16 to 22% polyethylene glycol (PEG) 4000 in 50 mM-citrate/phosphate buffer at pH 6·6 to 7·2. All buffers contained a trace (0·02%) of sodium azide and 10^{-4} M-DTT. Crystallization in the presence of cofactor was done under the same conditions but 0·3 mM-NAD^+ was added to all buffer solutions. Drops of 10 μl were used, consisting of 5 μl protein solution and 5 μl buffer with precipitant. Larger crystals were grown using the sitting-drop method, where drop sizes of up to 50 μl were used. Initial trials produced small (0·05 mm × 0·2 mm × 0·1 mm), monoclinic plates with approximate cell dimensions $a = 77$ Å, $b = 55$ Å, $c = 150$ Å, $\beta = 104°$, which not only diffracted weakly but also generally proved to be twinned. Further, the unit cell dimensions varied from batch to batch, although the cell volume remained constant. Addition of DTT to the buffer produced crystals only in the absence of NAD^+, but with a chunkier morphology and dimensions up to 0·4 mm. More typically the crystals were 0·2 mm × 0·2 mm × 0·2 mm and this has hampered photographic characterization in our laboratory.

A native data set was collected from the second form on station 9·6 at the Synchrotron Radiation Source, SERC Daresbury Laboratory, using the FAST area detector (Enraf-Nonius, Delft). Data were collected as 0·1° images over a range of 180° at a wavelength of 0·89 Å and processed using the MADNES program (Messerschmidt & Pflugrath, 1987). The auto-indexing facility of this program was used to search for the unknown unit cell. Refinement gave a monoclinic unit cell with dimensions, $a = 81{\cdot}24(6)$ Å, $b = 55{\cdot}75(4)$ Å, $c = 109{\cdot}60(7)$ Å and $\beta = 94{\cdot}26(9)°$. Integrated reflections were obtained using Kabsch's (1988) profile fitting technique, scaled together and precession photographs for the *hk0* and *h0l* zones *inter alia* were simulated. These photographs confirmed that the cell was indeed monoclinic, space group $P2_1$; the cell dimensions are not consistent with a unit cell of higher symmetry. The unit cell has a volume of 495,021 $Å^3$. This gives a V_m of 2·21 $Å^3$/dalton with two dimers per asymmetric unit (Matthews, 1968), and corresponds to a solvent content for the crystal of 44%. Data have also been collected at a wavelength of 1·5418 Å as 0·25° images on a Xentronics X-1000 detector mounted on a rotating-anode generator and were processed and refined with XDS (Kabsch, 1988) using the same cell as above. Simulated precession pictures from these data confirmed the space group as being $P2_1$ and match

some recently taken, 10° precession photographs (20 h exposure, CuKα, 40 kV, 30 mA). However, both simulated and true precession photographs reveal systematic absences at low resolution. The data set collected at the SRS (Synchrotron Radiation Source) shows similar systematically weak reflections, which conform to $h+l$ = odd, a *B*-face centring. In the standard *b*-axis setting for the monoclinic system, no such centring is permitted but these systematically absent reflections are only observed at resolutions less than about 4·0 Å. In fact, the initial autoindexing in XDS on one occasion produced the smaller cell related to the true cell by rotation about the *b*-axis. It appears, then, that there is a nearly exact relationship between the two independent dimers in the asymmetric unit. Work with the self-rotation function in both cells together with the search for suitable heavy-atom derivatives is well under way.

The crystal structure of DADHase will permit molecular interpretation of how the enzyme achieves the same reaction as the mammalian protein but by an apparently different mechanism. It will also permit a more rational approach to the generation of specific mutants designed to probe the enzyme mechanism and metabolic flux, already under way in Edinburgh and elsewhere (Ribas de Pouplana *et al.*, 1990; Chen *et al.*, 1990, 1991).

We are grateful to Professor Neil Isaacs, Drs Simon Phillips and Andy Freer for use of Xentronics area detectors, and also Drs Miroslav Papiz, Pierre Rizkallah, Colin Groom, Paul Taylor and Mary Turner for help with data processing and many useful discussions. This work is a contribution from the Edinburgh Centre for Molecular Recognition and we are grateful to the Science and Engineering Research Council and British Council Acciones Integradas Programme for financial support.

References

Borras, T., Persson, B. & Jörnvall, H. (1989). Eye lens δ-crystallin relationship to the family of "long chain" alcohol/polyol dehydrogenases. Protein trimming and conservation of stable parts. *Biochemistry*, **28**, 6133–6139.

Chambers, G. K. (1988). The *Drosophila* alcohol dehydrogenase gene–enzyme system. *Advan. Genet.* **25**, 39–107.

Chen, Z., Lu, L., Shirley, M., Lee, W. R. & Chang, S. H. (1990). Site-directed mutagenesis of Gly-14 and two "critical" cysteinyl residues in *Drosophila* alcohol dehydrogenase. *Biochemistry*, **29**, 1112–1118.

Chen, Z., Lee, W. R. & Chang, S. H. (1991). Role of aspartic acid 38 in the co-factor specificity of *Drosophila* alcohol dehydrogenase. *Eur. J. Biochem.* **202**, 263–267.

Eklund, H. & Brändén, C.-I. (1987). Alcohol dehydrogenase. In *Biological Macromolecules and Assemblies*, vol. 3, *Active sites of enzymes* (Jurnak, F. A. & McPherson, A., eds), pp. 74–114, John Wiley & Sons, New York.

Ghosh, D., Weeks, C. M., Grochulski, P., Duax, W. L., Erman, M., Rimsay, R. L. & Orr, J. C. (1991). Three-dimensional structure of a holo 3α,20β-hydroxysteroid dehydrogenase: a member of a short chain dehydrogenase family. *Proc. Nat. Acad. Sci., U.S.A.* **88**, 10064–10068.

Inoue, T., Sunagawa, M., Mori, A., Imai, C., Fukuda, M., Takagi, M. & Yano, K. (1989). Cloning and sequencing of the gene encoding the 72-kilodalton dehydrogenase subunit of alcohol dehydrogenase from *Acetobacter aceti*. *J. Bacteriol.* **171**, 3115–3122.

Jörnvall, H., Persson, M. & Jeffery, J. (1981). Alcohol and polyol dehydrogenases are both divided into two protein types, and structural properties cross-relate the different enzyme activities within each type. *Proc. Nat. Acad. Sci., U.S.A.* **78(7)**, 4226–4230.

Jörnvall, H., Persson, B., Krook, M. & Kaiser, R. (1990). Alcohol dehydrogenases. *Biochem. Soc. Trans.* **18**, 169–171.

Juan, E. & González-Duarte, R. (1980). Purification and enzyme stability of alcohol dehydrogenase from *Drosophila simulans*, *Drosophila virilus* and *Drosophila melanogaster* adh^s. *Biochem. J.* **189**, 105–110.

Kabsch, W. (1988). Evaluation of single-crystal X-ray diffraction data from a position sensitive detector. *J. Appl. Crystallogr.* **21**, 916–924.

Matthews, B. W. (1968). Solvent content of protein crystals. *J. Mol. Biol.* **33**, 491–497.

Messerschmidt, A. & Pflugrath, J. W. (1987). Crystal orientation and X-ray pattern prediction routines for urea-detector diffractometer systems in macromolecular crystallography. *J. Appl. Crystallogr.* **20**, 306–315.

Persson, B., Krook, M. & Jörnvall, H. (1991). Characteristics of short chain alcohol dehydrogenases and related enzymes. *Eur. J. Biochem.* **200**, 537–543.

Ribas, de Pouplana, Ll, Atrian, S., González-Duarte, R., Fothergill-Gilmore, L. A., Kelly, S. M. & Price, N. C. (1991). Structural properties of long and short-chain alcohol dehydrogenases: contribution of NAD^+ to stability. *Biochem. J.* **276**, 433–438.

Scopes, R. K. (1983). An iron-activated alcohol dehydrogenase. *FEBS Letters*, **156**, 303–306.

Thatcher, D. R. & Sawyer, L. (1980). Secondary-structure prediction from the sequence of *Drosophila melanogaster* (fruitfly) alcohol dehydrogenase. *Biochem. J.* **187**, 884–886.

van Deldon, W. (1982). The alcohol dehydrogenase polymorphism in *Drosophila melanogaster*; selection at an enzyme locus. In *Evolutionary Biology*, vol. 5 (Hecht, K. M., Wallace, B. & Prince, G. T., eds), pp. 187–222, Plenum Press, New York.

Villaroya, A., Juan, E., Egestad, B. & Jörnvall, H. (1989). The primary structure of alcohol dehydrogenase from *Drosophila lebanonensis*. *Eur. J. Biochem.* **180**, 191–197.

Winberg, J.-O., Hovik, R., McKinley-McKee, J. S. & Gonzalez-Duarte, R. (1986). Biochemical properties of alcohol dehydrogenase from *Drosophila lebanonensis*. *Biochem. J.* **235**, 481–490.

Edited by A. Klug